HUMAN PHYSIOLOGY

HUMAN PHYSIOLOGY

BY

F. R. WINTON, M.D., D.Sc.
Emeritus Professor of Pharmacology, University College London

AND

L. E. BAYLISS, Ph.,D.
Honorary Research Associate in Physiology,
University College London

FIFTH EDITION

WITH 228 ILLUSTRATIONS

J. & A. CHURCHILL LTD.
104 GLOUCESTER PLACE, LONDON, W.1.
1962

First Edition	. .	1930
,, ,,	*Reprinted*	1932
Second Edition	. .	1935
,, ,,	*Reprinted*	1936
Third Edition	. .	1948
,, ,,	*Reprinted*	1949
Fourth Edition	. .	1955
Fifth Edition	. .	1962

1841.

Printed in Great Britain

PREFACE TO THE FIFTH EDITION

PHYSIOLOGISTS continue to expand their knowledge of the subject ; so also in neighbouring fields do medical researchers, experimental zoologists, biophysicists and biochemists. What fails to expand, however, is the assimilating power of students. An introductory text-book must give them some general view and perspective of the whole field, including recent advances. Many experiments, happily, become replaced by less complicated experiments as new techniques make more direct evidence available. Some theories become modified and aggregated into a single more embracing theory. Nevertheless, our attempt to prevent expansion of this book has not quite succeeded, despite omission of much material that no longer seems quite as vital as it once did. The number of pages has, alas, increased by about 5 per cent. in the last seven years—moderate, perhaps, as inflation goes—and in part due to less use of " small print."

We welcome in this edition two new colleagues in our team of collaborators: Dr. G. I. M. Swyer and Dr. P. A. Merton. The chapter on Reproduction and the three chapters on Nerve and the Nervous System have accordingly been entirely re-written. The chapters on Respiration, Body Fluids, Urine, Muscles and Hearing have been largely re-written, and all the others subjected to thorough revision. We are grateful to our colleagues, as listed on p. x, who from first-hand knowledge of their subjects have made this possible. The number of separate chapters has been increased by fission of some of those in the previous edition. Notably, that on the Sense Organs has been divided into its component parts ; those properties and functions of the blood other than the carriage of oxygen and carbon dioxide have either been incorporated in the chapter on Body Fluids or put in a separate chapter, and Temperature Regulation has been given a chapter to itself.

This edition contains many new illustrations : some of these have been specially prepared, while others have been taken unchanged from original publications. For all these, whether modified or unchanged, we would like to record our thanks to the authors concerned and our gratitude for the cordial co-operation of the editors of the following journals : *Acta physiologica Scandinavica ; American Journal of Physiology ; Brain ; British Journal of Anæsthesia ; British Medical Bulletin ; Bulletin of the Johns Hopkins Hospital ; Clinical Science ; Nature ; Journal of the Acoustical Society of America ; Journal of the American Medical Association ; Journal of Biophysical and Biochemical Cytology ; Journal of General Physiology ; Journal of Neurophysiology ; Journal of Physiology* and the *Proceedings of the Royal Society.*

The D. Appleton-Century Co. ; Edward Arnold (Publishers) Ltd. ; Baillière, Tindall & Cox ; G. Bell & Sons ; Cambridge University Press ; W. Heffer & Sons Ltd. ; Hinrichsen Editions ; Longmans, Green & Co. Ltd. ; Macmillan & Co. Ltd. ; The Macmillan Co., New York ; Oxford

University Press ; The W. B. Saunders Co. ; Shaw & Sons ; Charles C. Thomas ; and The Williams and Wilkins Co. have all been good enough to grant permission for the use of certain illustrations from their publications and we thank them for this facility.

Finally, we must again express our appreciation of the courtesy and patience maintained by our publishers, and particularly of the skill and consideration shown by their artist, by Mr. J. Rivers and by Mr. A. S. Knightley.

F. R. WINTON
L. E. BAYLISS

LONDON
July, 1962.

EXTRACT FROM PREFACE TO FIRST EDITION

THE science of physiology has arisen as the confluence of a number of independent streams of inquiry. Chief among them, the problems of human disease called for a survey of the normal functions of the human body, and this inspired the development of experimental mammalian physiology and supplementary observations on the frog when the choice of a cold-blooded animal rendered the technique simpler. Hardly less in importance, the developments of chemistry in the direction of biochemistry, and of zoology in that of comparative physiology, have intermingled with the classical physiology, and the composite science of general physiology is now concerned more with the furtherance of natural knowledge than with applications to medicine. So it happens that the proper study of physiology demands an extensive knowledge of other sciences, and it is the more arduous since so young a science of necessity enjoys but few generalisations to co-ordinate the bewildering diversity of its biological material.

As the scope of physiology has expanded, the attempt to acquaint the medical student with the salient facts has trespassed overmuch on his opportunities for leisurely reflection ; any real assimilation of the matter presented to him has become increasingly difficult, and much of the subject is forgotten soon after the urgent need for its retention is passed. Hence the complaints of teachers of medicine that their students remember most of the relevant facts of human physiology so tenuously that they fail to apply them to clinical problems.

The medical student is expected to know too much and to think too little. He has two chief needs, a training in scientific methods and a knowledge of the properties of the human body. They can, it seems, best be satisfied, within the comparatively short period of his education available for physiology, by confining his attention to " Human Physiology," and eliminating from his curriculum those parts of physiology which have no immediate bearing on the happenings in the body of man.

If, then, it be convenient to distinguish Human Physiology as a separate subject for teaching purposes (and we are not here concerned with the organisation of research), it would clearly be desirable to base the subject on direct observations on man, and to proceed by way of analysis of these to experiments on animals directed to the solution of the problems raised by those initial observations. This arrangement often reverses the chronological sequence of discovery, and can only be adopted without loss of logical coherence in suitably developed parts of the subject. Muscular activity, for example, lends itself to such treatment, whereas the secretion of urine in man is still too little understood to make it a satisfactory approach to the systematic description of the functions of the kidney. Indeed, it is only the recent rapid progress in certain fields that has made it feasible now to attempt

their presentation as an analysis of the properties of man instead of a synthesis of experiments on animals and isolated organs.

The book has been kept short by omissions rather than by compression. Much anatomy, histology and biochemistry that is commonly included in text-books of physiology is now so admirably treated in special books on these subjects that duplication here has been deemed unnecessary, and only an outline of the immediate facts needed to complete the description or argument under consideration has been included.

F. R. WINTON
L. E. BAYLISS

LONDON
September, 1930.

We are grateful to the following colleagues for writing certain chapters, as indicated in the Table of Contents.

H. Barcroft, M.A., M.D., F.R.S.
Professor of Physiology, St. Thomas' Hospital Medical School, London S.E.1.

M. de Burgh Daly, M.A., M.B., B.CH.
Professor of Physiology, St. Bartholomew's Hospital Medical College, London E.C.1.

R. A. Gregory, M.SC., D.SC., PH.D., M.R.C.S., L.R.C.P.
Professor of Physiology, The University, Liverpool 3.

D. H. Smyth, M.D., M.SC.
Professor of Physiology, The University, Sheffield 10.

G. I. M. Swyer, D.PHIL., D.M., M.R.C.P.
Consultant, Obstetric Hospital, University College Hospital, and Lecturer, U.C.H. Medical School, London W.C.1.

G. W. Harris, M.A., SC.D., F.R.S.
Professor of Anatomy, University of Oxford.

P. A. Merton, M.B.
Lecturer in Physiology, University of Cambridge.

D. R. Wilkie, M.R.C.P., M.D., B.S.
Reader in Experimental Physiology, University College, London W.C.1.

J. A. B. Gray, M.A., M.B., B.CHIR.
Professor of Physiology, University College, London W.C.1.

M. G. Eggleton, D.SC., M.R.C.S., L.R.C.P.
Reader in Physiology, University College, London W.C.1.

CONTENTS

PROLOGUE

The dead and the living body differ. How they differ and how the living body sees, moves, digests, keeps warm, and so on, is the province of physiology. We see a man lift an arm or walk without staggering and as physiologists we wonder what is going on in his muscles and nervous system. To discover things of this kind, we have first to take the machine to pieces, much as we should have to take a motor car to pieces to explain its varied performance on the road. You cannot, of course, take a man to pieces, so the corresponding tissues of animals, say the limb muscles, intestine or heart, have to be examined one by one after removal from animals immediately after death. If the property of an organ is found to be much the same in a number of animals, such as the frog, rabbit, cat, dog and monkey, we presume that it may well be much the same in man. This basic knowledge about individual organs can be extended to a study of their functions and interactions in anæsthetised animals. Only then, as a rule, can painless methods be devised for similar and further studies on unanæsthetised animals and man. Thus physiology extends from physical and chemical processes in cells and tissues to elaborate performances of whole animals, but it does not extend very far into studies of behaviour or mental affairs which are the province of psychology.

Physiology has another boundary ; it is primarily concerned with normal animals. Disease processes belong to the subject matter of pathology and medicine. Practising physicians, however, need a vivid awareness of physiological processes in the body for both diagnosis and treatment. Normally they lack the leisure to submit disorders presented by individual patients to extensive physiological analysis ; moreover, the treatment of the sick cannot be delayed till the outcome of such analysis is known. Treatment depends on speedy diagnosis which is based on the symptoms reported by the patient, on physical signs observed by the physician and perhaps on a few simple tests such as X-ray, histological or biochemical examinations. Such slender elements of evidence must be fitted into the whole relevant knowledge of physiological processes in the body before they have much value beyond rule-of-thumb indicators, yielding "slot-machine diagnosis." So it happens that important advances in physiology are made in the research departments of medicine, particularly in fields somewhat neglected by physiologists who may be preoccupied with studies without such applications to medicine. Certain diseases such as diabetes mellitus have, however, been extensively studied by physiologists sometimes with dramatic improvements in treatment as a consequence.

The most noticeable and measurable sign of deterioration in ill-health is often the reduction of the amount of muscular exercise that can be endured without excessive breathlessness and thumping of the

heart. If a sedentary clerk and an athlete run side by side to catch a train, the first may arrive panting, sweating and exhausted and take quite some time to recover, while the second may suffer little discomfort and that soon over. Neither would be regarded as ill unless his performance deteriorated to the point of interfering with his normal occupation. A third person may find that, for example, climbing slowly up a staircase may engender such breathlessness as to demand a rest on reaching each floor, yet the physiological mechanisms controlling breathing are much the same in all three. Thus, physiology is essential to the understanding of both normal and disease processes.

The line dividing the living and the dead is not easy to draw. If you prod a live animal it is apt to move, but neither anæsthetised animals nor most plants move actively when disturbed. Growth and reproduction are regarded as characteristic of the living state, but many adults past growth and beyond child-bearing are far from dead and still appear irritable to prodding.

Men and large animals are certainly dead when the heart has stopped beating for so long that there is no hope of recovery. In them, the blood pumped by the heart is essential for carrying foodstuffs and their products in the intestine to organs like muscles, to which, also, oxygen is carried by blood from the lungs. Energy derived from chemical reactions resembling burning can then be converted into mechanical or other forms of useful energy, while the waste products of the combustion are carried away, again by the blood, the carbon dioxide to the lungs and most soluble products to the kidneys, where they are eliminated from the body. Perhaps the continued use of oxygen as part source of chemical energy is one of the more widespread features of life, but anaerobic organisms provide exceptions ; moreover, power for industry and traffic is still derived from fuels oxidised in apparatus which is indubitably dead.

In animals of microscopic size, in contrast to larger animals, blood circulation is not needed. Every part is so near the surface of the cell or organism that it can derive oxygen and foodstuffs by direct diffusion from the surface. A blood circulation is, however, essential in larger animals and this may be illustrated by some features of muscular exercise in man.

Muscle occupies much of the total body weight and when active, each gram of muscle uses oxygen and produces waste products more rapidly than any other tissue. The large output of carbon dioxide itself is one of the factors producing the well-known increase in breathing during strenous exercise, and this also secures the increase in oxygen intake. A large increase in blood flow is needed to carry the extra carbon dioxide and oxygen to and from the lungs and is produced by the heart which accordingly beats at a higher rate and puts out more blood per beat. Moreover, muscles resemble man-made engines in being unable to convert more than, at best, about one quarter of the chemical energy into mechanical work. Three-quarters or more, therefore, of the chemical energy is converted to heat. Again, the blood

is essential for transporting the extra heat from muscles to the skin from which the heat can be lost by radiation and conduction. Indeed, in healthy people the severity of the exercise they can take is generally limited, not by their muscles, but by the rate at which the heart can pump blood round the body.

Cardiac output can be measured in man, but not easily enough to be a useful test in clinical practice. Heart-rate, easily measured as the pulse of the radial artery at the wrist, is a most useful quick guide to change in cardiac output, but when the frequency reaches about 180/minute the time available for the heart to refill between contractions becomes too short and the output per beat falls. Below this limit, pulse-rate can be used as a guide to the degree of physical fitness in two ways. First, during moderate exertion the pulse-rate will rise more in sick people than would be expected in normally fit people. Secondly, during rest following exertion, the pulse-rate will return to normal more slowly.

Undue fatigue is well known to accompany ill-health. Fatigue is a state which may reduce or end longer-lasting muscular exercise or many other forms of activity. It is said to supervene when an activity, which has been sustained for some time, diminishes although the incentive to maintain it remains unchanged and effective, the standard of performance being completely restored after a period of rest. Formerly, fatigue was attributed to the accumulation in the body of metabolic products of activity known as "fatigue products." Though this may happen in a few kinds of activity, it is now considered incapable of explaining most forms of fatigue. Different kinds of activity may be impaired by different processes which may concern primarily almost any of the physiological systems in the body.

An isolated frog muscle contracts less and less in response to repeated electrical stimulation of constant strength : if given a rest, the contractions regain their original size. By analogy, it is said to have become "fatigued." This is attributed to inadequate supply of oxygen diffusing from the surface of the muscle to the fibres and leading to an accumulation of lactic acid. Accordingly, it is difficult to fatigue a single muscle in an anæsthetised mammal by repeated electrical stimulation because it is well-provided with oxygen from its own blood circulation. In very severe exercise in man involving many muscles, however, the oxygen supply to muscles is restricted by the cardiac output, as mentioned earlier, and lactic acid correspondingly accumulates in the body and reduces muscular activity. In less severe exercise the lactic acid is oxidised to carbon dioxide and water ; but if such exercise continues till the animal, say a dog running on a treadmill, is exhausted, the fatigue is due to lack of fuel. If glucose is administered periodically, the dog can continue to run for much longer without exhaustion.

Another factor which produces fatigue after even short spells of moderate exercise is the undue rise in body temperature which occurs in a hot moist atmosphere, or if evaporation is reduced by wearing a rubber or oilskin coat. Reduction in blood volume which reduces

blood circulation in the brain often makes people feel tired and " go slow " accordingly. When men stand strictly to attention for half an hour or so, they may even faint. The loss in blood volume is here due to high blood pressure in the veins and capillaries of the immobile legs with outward filtration of water into the tissue fluid. There are many ways other than blood or oxygen lack in which the brain can become fatigued, mental work among them as some students know from experience.

The activity, and even survival, of the cells and tissues of an animal or man depend on the properties of the fluids surrounding them. For example, chemical processes in cells are catalysed by enzymes, at varying rates according to temperature and acidity ; but some vary more than others. Particularly in the more complex species, therefore, body temperature, the acidity and osmolar concentration of body fluids and many other factors must remain nearly constant if physiological processes are to continue in normal balance. Such ideas were crystallised by Claude Bernard (1870) ; the higher animals were said to have an *internal environment* which is maintained remarkably constant, even during great changes of the external environment, such as climate, the amount and kinds of food and water taken in, and the extent of muscular activity. The pattern of regulatory processes in the body which operate to control the internal environment within so narrow a range of variation was called *physiological homeostasis* by W. B. Cannon (1929) and is one of the essential factors in maintenance of health and, indeed, in survival.

The following chapters include many descriptions of such regulatory processes appertaining to the physiological systems concerned. The interactions between these processes, to achieve homeostasis, will be considered further in the epilogue following the chapters devoted to the individual physiological systems.

CHAPTER 1

CIRCULATION OF THE BLOOD—THE HEART

We owe to Harvey (1628) the conception and proof of the idea that the blood circulates, and as this step marks the beginning of modern physiology, it is of more than usual interest to note his argument. Harvey was able to show that the valves in the heart are so arranged as to allow the passage of blood in only one direction. Further, by watching the motion of the heart in the living animal he concluded that blood is expelled from the ventricles into the pulmonary artery and aorta during systole, and enters the heart from the venæ cavæ and pulmonary veins during diastole. He calculated that if only a drachm of blood were expelled at each beat, the heart would in half an hour use up all the blood in the body, and so empty the veins completely and distend the arteries ; from this he concluded that the blood must move in a circle, entering the veins from the arteries. Proof that the blood flows continuously in one direction he found in the cutaneous veins of the human arm (Fig. 1.1). It had been shown previously by Fabricius (1603) that if a ligature is tied around the arm the veins swell up distally, and present along their course little swellings which mark the position of valves allowing the passage of blood in only one direction, towards the heart. The significance of this discovery was lost on Fabricius, but was appreciated by Harvey, who made a further observation, which anyone may repeat on his own arm. The middle finger of the left hand is pressed firmly on a prominent cutaneous vein of the right forearm, and blood is massaged out of the vein proximally by rubbing the forefinger firmly up the vein past the next valve. If the forefinger is lifted, the vein fills from above only as far as the valve, but when the middle finger is lifted the whole stretch of collapsed vein rapidly fills from below. This process may be repeated indefinitely, blood always entering from the periphery ; we must conclude with Harvey that blood is always entering the veins from the distal side, and the only source of this blood is the arterial system.

The missing link in Harvey's argument, namely, the connection between the arteries and veins, was supplied forty years later by Malpighi, when he discovered the capillaries.

The general arrangement of the circulation is shown in Fig. 1. 2. In man, as in all mammals, the circulation consists of two circuits connected in series, the greater or systemic, and the lesser or pulmonary circuit ; accordingly the heart consists of two pumps, one for the lungs and the other for the rest of the body. Blood is pumped by the right heart into the pulmonary artery and on through the pulmonary capillaries to the pulmonary veins, and so enters the left heart. The left heart ejects the blood into the aorta, and then into the branching

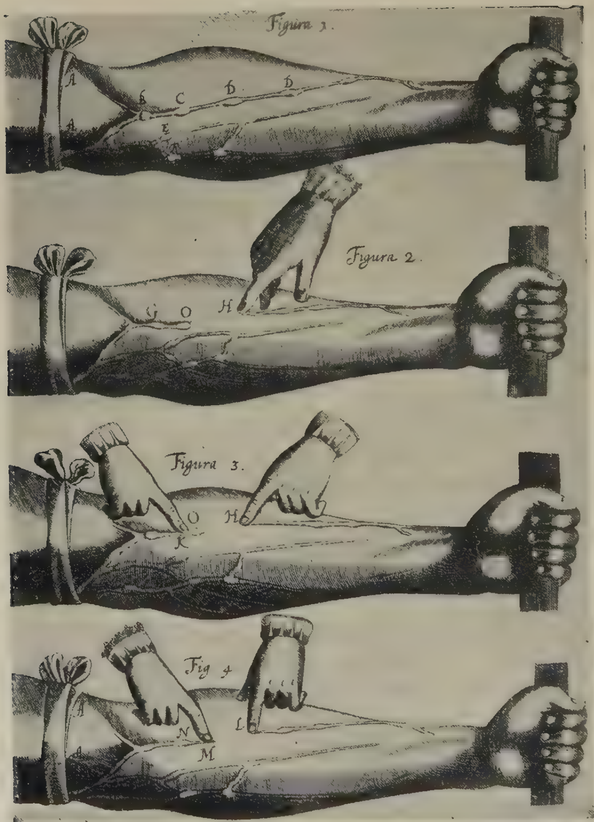


FIG. 1. 1. Harvey's Figures illustrating the **Unidirectional Flow of Blood** in the Veins.

(1) The bandage A A is tied round the arm above the elbow, constricting the veins which become distended. The position of each valve is indicated by a swelling or knot, B, C, D, E, F. (2) The blood is pressed out of a vein from H to O with one finger, while another keeps the vein closed at H. No blood runs back past the valve at O. (3) If the vein is pressed by another finger at K, no blood can be forced backwards past the valve at O. (4) The vein is closed with one finger at L, as before, and emptied by stroking with another finger, M, towards the valve at N ; the vein continues empty until the finger at L is removed, when it rapidly fills up from the periphery.

arteries which distribute it to the various organs of the body. From the capillaries of these organs, the blood is collected into veins, and then returns to the right heart *via* the inferior and superior venæ cavæ. During its passage through the pulmonary capillaries, the blood gains oxygen from the air in the lungs, and loses carbon dioxide : in the capillaries of the other organs, supplied from the left heart, the blood loses oxygen to the tissues, and gains carbon dioxide from them.

The blood flow to the various organs and tissues depends on the total output of the heart per minute, and on the proportion of this output

which is sent to each of them. The regulation of the cardiac output, and of the eventual distribution of the blood, are thus the central problems of the physiology of the circulation.

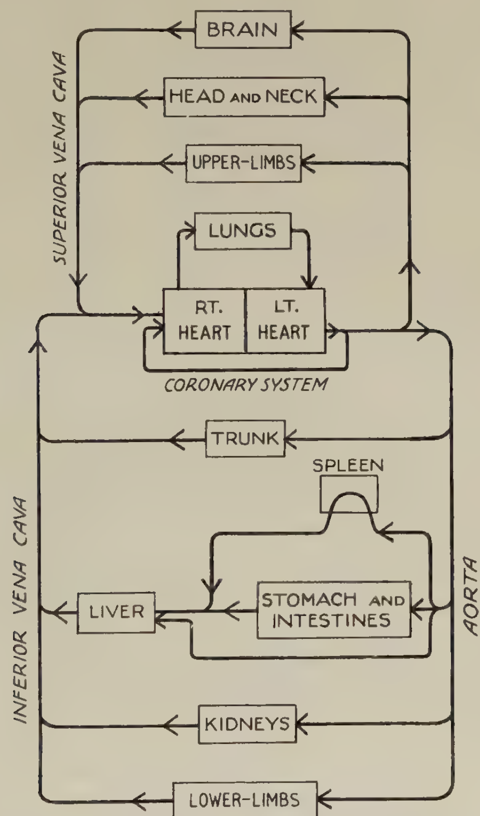


FIG. 1. 2. Scheme of the **Human Circulation.**

The Heart

The heart is divided longitudinally into the right and left hearts, each consisting of two communicating chambers, the atrium (or auricle) and ventricle. The capacity of each ventricle when fully relaxed is about 140 to 200 ml. in man. The heart wall consists essentially of muscle (the myocardium), which has an inner covering (the endocardium), lined by endothelium, and an outer covering (the epicardium or visceral layer of the pericardium). Covering the heart is a fibrous sac, the pericardium, which at its attachment to the great vessels is reflected over the outer surface of the heart, thus leaving between its outer or parietal layer and its inner or visceral layer a potential space, the pericardial cavity. The pericardium is attached to the surrounding structures, and thus partially fixes the heart while it allows it such freedom of movement as is essential for its contraction.

The heart muscle consists of quadrilateral cells, which are joined longitudinally to form fibres and anastomose with neighbouring cells by short bridges. It is an arrangement in which the cells communicate with one another and is termed a syncytium. The properties which these muscle cells possess in common with other contractile tissues will be dealt with in the chapter on Muscle (Chapter 16). Over the auricles the muscular wall is relatively thin, over the ventricles relatively thick; the wall of the left ventricle is four times as thick as that of the right. The thickness of the muscular wall of each chamber thus

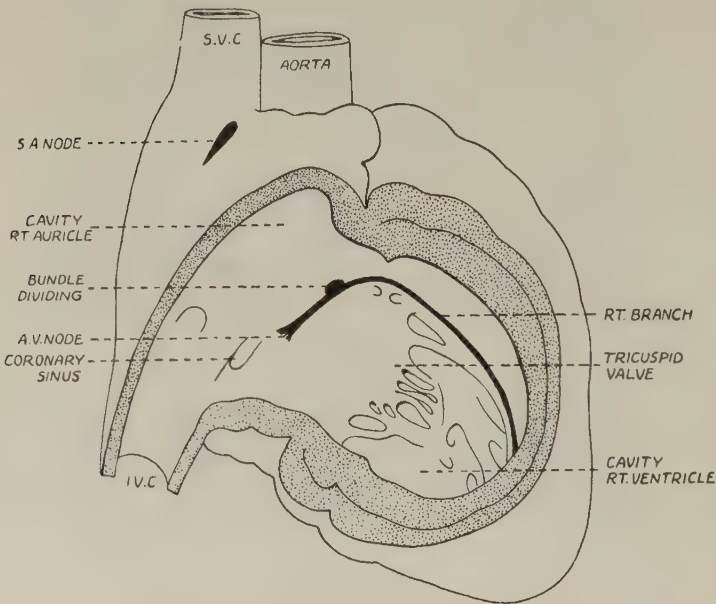


Fig. 1. 3. A diagram of the human heart to show the **Sino-auricular Node** (S.A. node), the **Auriculo-ventricular Node** (A.V. node) and **Bundle of His.** The walls of the inferior vena cava, right auricle and right ventricle have been partially removed to expose the septa. The cut surfaces are stippled.

corresponds to the tension developed during its contraction. The muscle fibres of the right auricle are continuous with those of the left, those of the right ventricle are continuous with those of the left ventricle. The muscle fibres of the auricles, however, are separated from those of the ventricles by a fibro-tendinous ring, the auriculo-ventricular ring. The heart muscle is modified to form two important structures (Fig. 1. 3). The first or **sino-auricular node** lies close to the junction of the superior vena cava with the right auricle and is about 2 cm. long and 2 mm. wide in man. It consists of a plexus of fine muscle fibres embedded in fibrous tissue. The second is the **auriculo-ventricular connection,** which forms the only functional junction

between the muscular tissues of the auricles and the ventricles. This begins at the base of the inter-auricular septum close to the mouth of the coronary sinus as the auriculo-ventricular node, composed of slender interlacing muscle fibres. Continuous with the auriculo-ventricular node is the auriculo-ventricular bundle (of His), which runs across the fibrous ring between auricles and ventricles and enters the inter-ventricular septum, where it divides into right and left branches distributed to the appropriate ventricles. Each branch is continuous with a network of large, poorly striated cells, rich in glycogen, the Purkinje tissue, which forms a plexus under the endocardium of each ventricle.

The cavity of each auricle is separated from that of the corresponding ventricle by an auriculo-ventricular valve, a fibrous membrane covered with endocardium and arising from the auriculo-ventricular ring. On the right side the valve is divided into three flaps (tricuspid), on the left into two (mitral). To the ventricular aspects of the margins of these valves are attached tendinous chords (chordæ tendineæ), which terminate in nipple-like projections of the ventricular muscle (papillary muscles). These valves are so arranged that when blood flows from auricle to ventricle the valves lie flat against the ventricular wall. When the ventricular pressure rises above the auricular pressure the valves are floated out by eddies and seal the auriculo-ventricular openings ; the chordæ tendineæ, aided by the contraction of the papillary muscles, prevent the valves from being thrust out into the auricular cavity. The openings of the right ventricle into the pulmonary artery and of the left ventricle into the aorta are each guarded by semilunar valves, consisting of three semi-circular pockets whose cavities face away from the ventricles. The openings of the caval veins into the right, and of the pulmonary veins into the left auricle are unguarded ; they are, however, sealed at the beginning of auricular systole by the contraction of the auricular muscle fibres surrounding them.

The Cardiac Cycle. The contraction of the heart can be seen in man by means of X-rays, and in experimental animals directly by opening the chest (Fig. 1. 4). The cardiac cycle begins with a simultaneous contraction of both auricles (auricular systole), which are seen to become paler and smaller in size. After a short pause both ventricles contract (ventricular systole), at first becoming paler and more rounded, then smaller in size. As the ventricles empty, the aorta and pulmonary artery fill. After contraction, each chamber relaxes (diastole) and then gradually fills, to empty again at its next beat. The hardening and change in shape which constitute the first phase of ventricular contraction are accompanied by a thrusting of the apical region of the ventricles against the chest wall. This thrust commonly moves the overlying intercostal space, and the movement, known as the cardiac impulse (" apex beat "), indicates the point at which the region of the apex of the heart lies. The position of the impulse has great importance clinically in indicating the size and position of the heart.

In physiological terminology the terms " systole " and " diastole " are applied to each chamber of the heart. In man the sequence of

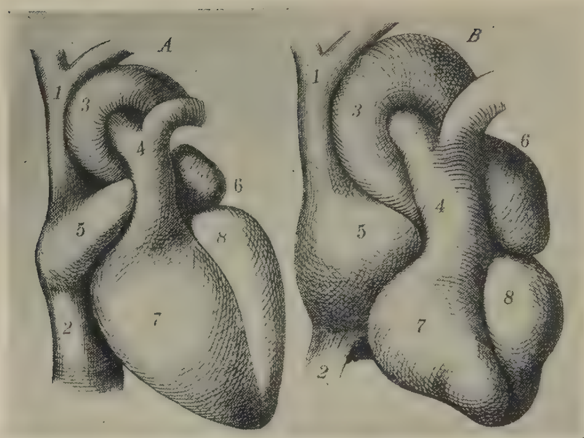


FIG. 1. 4. The General Aspect of the **Mammalian Heart** in Auricular Systole (A) and at the end of Ventricular Systole (B).

1, superior vena cava ; 2, inferior vena cava ; 3, aorta ; 4, pulmonary artery ; 5, right auricle ; 6, left auricle ; 7, right ventricle ; 8, left ventricle. (Modified from Rollett.)

contraction of auricles and ventricles can be recorded by the following graphic methods.

(1) **Mechanical Records—The Polygraph.** The ink polygraph records simultaneously on a strip of moving paper pulsations of the jugular vein and of the radial artery or heart's impulse. Chief interest in the records so obtained (Fig. 1. 5) attaches to those of the venous pulse. In normal subjects this consists of three waves in each cycle, *a*, *c*, and *v*. The *a* wave is due to auricular contraction and is caused by the arrest of venous inflow to the heart by constriction of the mouths of the great veins. The *c* wave indicates ventricular systole and is largely transmitted from the carotid artery. The *v* wave is of less importance and is largely due to slowing of the venous flow consequent on the filling of the heart in diastole. The object of the radial pulse tracing is to time ventricular systole and thus to identify the *c* wave of the jugular pulse ; the *c* wave of the jugular pulse occurs about 1/10th second before the radial pulse, the difference in time corresponding to the difference in their distances from the heart (see p. 29). The most important feature of the

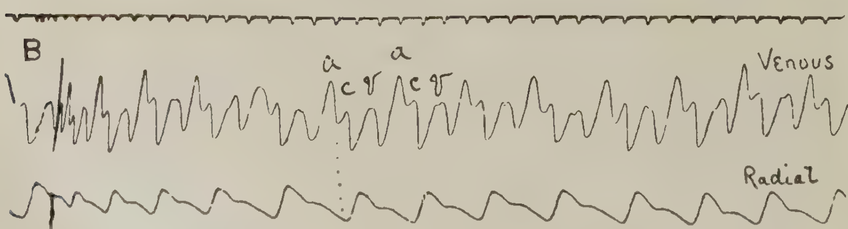


FIG. 1. 5. A simultaneous **record of the jugular and radial pulses** obtained by the polygraph ; the time marker records $\frac{1}{5}$ second. In the radial pulse curve the main upstroke is the primary wave ; the small hump following it is the dicrotic wave. (From Lewis' "Mechanism and Graphic Registration of the Heart-beat.")

jugular pulse is the *a* wave. In the early stages of heart block where conduction through the bundle of His is impaired, the interval between auricular and ventricular contractions and that between the *a* and *c* waves is prolonged. In the later stages, the ventricle ceases to respond to each auricular systole, and the *a* waves are not always followed by *c* waves. Lastly, in auricular fibrillation, where co-ordinated contraction of the auricles has ceased, the *a* wave is absent.

(2) **Electrical Changes.** The contractions of the heart, like those of other muscular organs, are associated with changes in electrical potential, and if the subject is at rest the quick changes of potential which may be led off from a pair of limbs are almost entirely cardiac in origin. After amplification, these may be displayed on an oscilloscope and recorded photographically, or recorded directly on paper by a suitable pen-recorder.

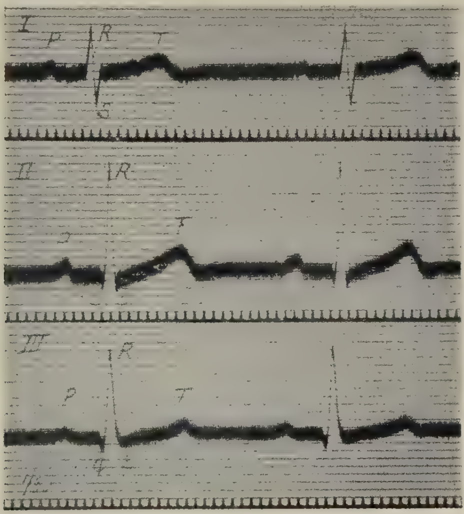


FIG. 1. 6. Typical **Electrocardiograms.**

I. Lead *I*, right arm to left arm. II. Lead *II*, right arm to left leg. III. Lead *III*, left arm to left leg. (From Lewis' "Mechanism and Graphic Registration of the Heart-beat.")

In normal subjects the record (electrocardiogram) has the form shown in Fig. 1. 6. Observation and experiment has shown that the P wave is due to the auricular contraction, the QRST complex to ventricular systole. Such records give us accurate information as to the spread of the cardiac excitation wave and have helped greatly to elucidate the nature of many irregularities of the heart's action.

The Origin and Conduction of the Heart-beat

If a mammalian heart is excised, it will continue to beat for hours, provided that an adequate supply of warm oxygenated fluid of suitable composition (Chapter 8, p. 244) is supplied to the muscle through the coronary vessels from the aorta. It is clear, then, that the origin of the heart-beat is independent of any connection with the rest of the body. Now the heart contains nerve ganglia, chiefly derived from the vagus,

but even if these are dissected out in the cold-blooded heart, the beat continues ; strips of auricular and ventricular mammalian muscle devoid of ganglia may contract rhythmically if placed in warm oxygenated Ringer-Locke solution. Further, in the chick embryo the heart begins to beat before it has received any nerves. Thus we may conclude that the beat originates in the heart muscle itself.

Although heart muscle is thus endowed with the property of contracting rhythmically, different parts of the heart behave differently in this respect. This is most easily shown in the classical experiments of Stannius on the heart of the frog (Fig. 1. 7). If a ligature is tied tightly around the junction of the sinus venosus with the auricles, the auricles and ventricle stop beating while the sinus continues at the same rate as before. After five to thirty minutes the detached part of the

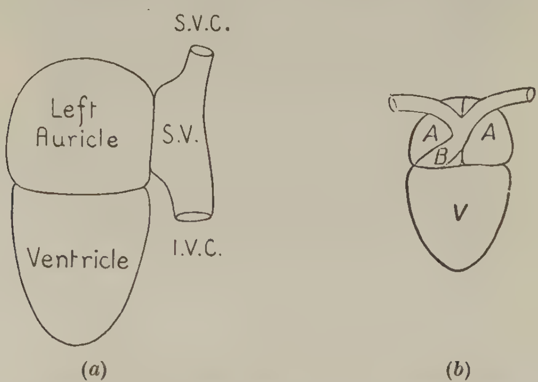


Fig. 1. 7. **Diagram of the frog's heart** (*a*) from the side, and (*b*) from the ventral aspect. The first Stannius ligature is tied tightly round the junction of the sinus venosus and right auricle (lying behind the left in the figure), the second between auricles and ventricle. S.V. = sinus venosus ; S.V.C. = superior vena cava and I.V.C. = inferior vena cava ; *B* = bulbus arteriosus.

heart begins to beat, but at a slower rate than that of the sinus ; the auricle contracts before the ventricle. If now a ligature is tied tightly between auricles and ventricle (second Stannius ligature), the auricles continue to beat as before, while the ventricle stops, after making a few rapid beats due to the stimulus of the ligature ; the ventricle begins to beat again after about an hour at a very slow rate. Thus although each chamber of the heart is able of itself to contract rhythmically, the frequency of contraction varies from the sinus venosus at one end to the ventricle at the other ; in the intact heart the rate of beating is that of the fastest chamber—the sinus venosus. This experiment thus suggests that in the frog the heart-beat begins in the sinus region and spreads over the chambers successively ; that the same is true of the mammalian heart will now be shown.

The Origin and Spread of the Electrical Changes in the Mammalian Heart. By placing electrodes, connected to a suitable recording instru-

ment, on various parts of the heart's surface, the electrical changes associated with the excitation of these parts can be observed and accurately timed. In this way, it was shown by Lewis that the electrical change which accompanies contraction of the mammalian heart begins in the sino-auricular node which lies in that part of the heart corresponding with the sinus venosus of the frog. From here the wave of electrical disturbance radiates in all directions over the auricular muscle, for the heart muscle cells are joined together as a syncytium, and for a short distance up the great veins. When the wave arrives at the auriculo-ventricular node it passes thence along the auriculo-ventricular bundle through the Purkinje tissue and into the ventricular muscle. The rate of conduction through the Purkinje tissue is very rapid (500 cm. per second) as compared with its rate through the ventricular muscle (50 cm. per second). Thus in spite of their size all parts of the ventricles contract almost simultaneously. That the normal heart-beat actually originates in the sino-auricular node (the pace-maker) is confirmed by the fact that when this structure alone is warmed or cooled, the heart-beat quickens or slows respectively ; when other parts of the heart are similarly warmed or cooled the frequency of the heart-beat is unaltered.

Heart Block. This occurs clinically in cases in which the bundle of His (Fig. 1. 3) is affected by some pathological process. The block may be either partial or complete. In partial heart block, most but not all of the auricular contractions excite ventricular ones. Often every third or fourth auricular beat fails to do so ; the radial pulse is then " regularly irregular." In complete heart block the auricles beat normally and the ventricles contract quite independently and much more slowly, thirty to forty times a minute. If complete heart block occurs suddenly, the ventricle may stop entirely for several seconds before taking up its independent rate of beat. During the period of complete ventricular quiescence, consciousness is lost, and an epileptiform seizure may occur, due to cessation of the circulation to the brain (Adams-Stokes attack).

Extra Systoles. These may occur in normal subjects and usually arise outside the pacemaker (ectopically).

Auricular Flutter and Fibrillation. These two forms of disordered heart action in man are characterised by extremely rapid auricular beats as judged by the waves of the electrocardiogram. In auricular fibrillation the auricular waves occur irregularly at the rate of about 400 to 500 per minute ; in flutter the waves are regular at 250 to 300 per minute. The bundle of His is incapable of conducting impulses at such rates and a variable degree of heart block is always present, the ventricle usually beating at about 100 to 150, perhaps regularly in flutter, always irregularly in fibrillation.

Conditions analogous to flutter and fibrillation can be produced experimentally by stimulating the auricle with rapid rhythmic electric shocks, or by touching it with a brush dipped in aconitine. At first sight, the fibrillating auricle appears to have stopped beating. However, on close inspection the whole surface is seen to be writhing and shimmering. Numerous small waves replace the P-wave of the electrocardiogram. According to Lewis, in flutter and fibrillation the excitation wave circulates continuously through the auricular muscle around the mouths of the caval veins. But results obtained more recently by means of the high-speed camera and by the cathode ray oscillograph, have thrown some doubt on this classical explanation of flutter and fibrillation ; the movements are not circus in character, but radiate from some fixed point in the auricle. This finding is supported by experiments on auricles made to fibrillate by applying aconitine. Cooling the spot to which

the aconitine was applied abolished the fibrillation and restored the normal rhythm.

Ventricular Fibrillation. This occurs in man after blockage of the coronary vessels and during electrocution. Owing to the cessation of ventricular contractions, it is rapidly fatal. It can be induced experimentally by faradic stimulation of the ventricles.

The Pressure Changes accompanying the Cardiac Cycle

Although the heart consists of two pumps, the right and left hearts, these work simultaneously and in the same way, and it will be convenient to describe the pressure changes only of the left auricle and ventricle ; those of the right auricle and ventricle are similar and simultaneous though of less magnitude. Of the various events of the cardiac cycle whose time relations are shown in Fig. 1. 8, the jugular pulse and electrocardiogram have been considered already ; the pressure changes in heart and aorta and the change in ventricular volume will now be described. The pressure changes have been recorded by electrical or optical manometers (p. 35) connected by means of rigid tubes filled with fluid to cannulæ thrust into chambers of the heart. The volume changes were recorded by means of the cardiometer to be described on p. 18.

The intra-auricular pressure curve shows waves of rise of pressure (positive waves) corresponding to auricular systole (first positive wave) and the sudden closure of the auriculo-ventricular valves (second positive wave). The pressure then abruptly falls as the relatively emptied auricle relaxes. Blood flows into the auricles from the great veins, producing a gradual rise of pressure (third positive wave), which is interrupted when the auriculo-ventricular valves open and put the relaxed and relatively emptied ventricle into connection with the auricle. It will be noted that during diastole both intra-auricular and intra-ventricular pressures are below atmospheric pressure, which is represented as O mm. in Fig. 1. 8. This is not due to any sucking action of the heart, but to the transmission of the " negative pressure " in the thorax through the slack heart wall.

Ventricular and Aortic Pressure. While the ventricles are quiescent, blood is flowing into them from the auricles, and the intra-ventricular pressure is slightly lower than, and closely follows, the intra-auricular pressure. With the onset of ventricular contraction the intra-ventricular pressure rises abruptly until it exceeds the aortic pressure, the aortic valves now open, blood is forced into the aorta, and the two pressures mount together. As ventricular ejection begins to decline the ventricular and the aortic pressures begin to fall, at first slowly, then rapidly, as the ventricle passes into diastole. The aortic valves now close ; the aortic pressure falls slowly as blood flows out at the periphery and the ventricular pressure falls abruptly. The ventricular pressure now falls below auricular pressure, which has been rising owing to the venous inflow, the auriculo-ventricular valves open and blood flows into the ventricle, gradually raising its pressure until the next cardiac cycle begins.

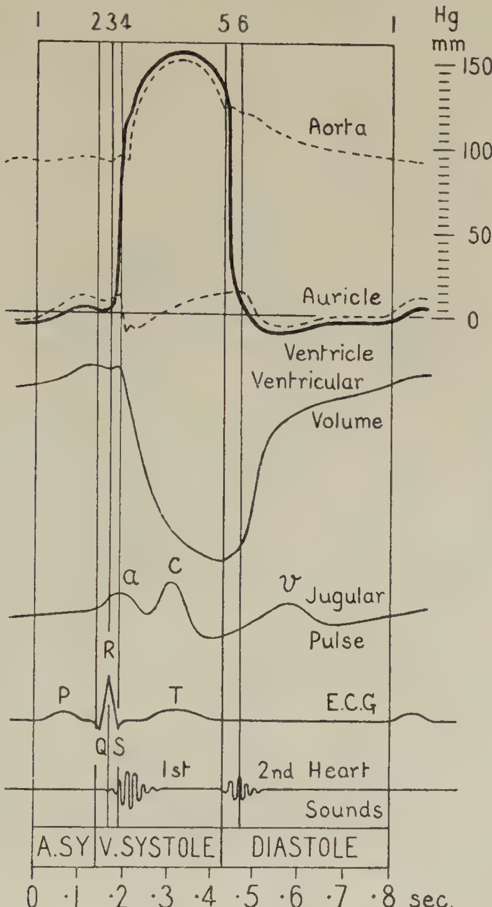


FIG. 1. 8. **The Sequence of Events in the Cardiac Cycle.**

The upper four curves have been taken from actual records obtained from the dog's heart (Wiggers) ; they represent the pressure changes in aorta (broken line), left ventricle (continuous line), and left auricle (broken line), and the curve of ventricular volume. The lower three curves, representing the jugular pulse, the electrocardiogram, and the heart sounds, have been reconstructed from data obtained on human subjects (Lewis). The vertical lines represent the following events : 1 = auricular excitation (P wave of electrocardiogram), 2 = ventricular excitation (Q wave of electrocardiogram), 3 = auriculo-ventricular valves close, 4 = aortic valve opens, 5 = aortic valve closes, 6 = auriculo-ventricular valves open. A.SY represents the duration of auricular systole. (More usual times for the duration of auricular systole in man would be 0·10 second, of ventricular systole 0·24 second, and of diastole 0·46 second.)

Ventricular Volume. The volume of the ventricles is slightly increased during auricular systole. The onset of ventricular contraction is associated with no diminution of volume, for the ventricle is now a closed cavity separated from the auricle by the auriculo-ventricular

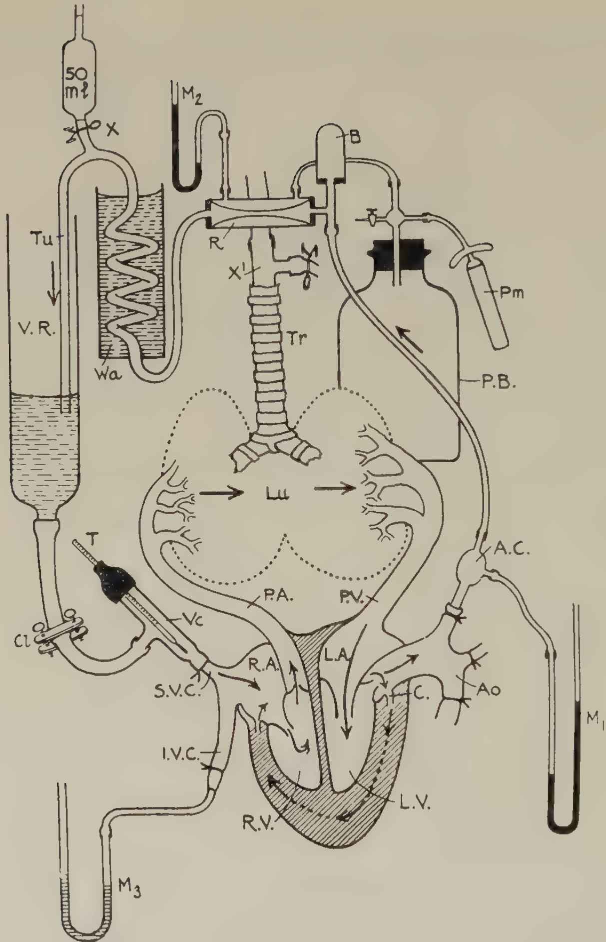


FIG. 1. 9. **The Heart-Lung Preparation.**

The direction of the flow of blood is shown by arrows. V.R., venous reservoir ; Cl, clamp for adjusting the rate of inflow to the heart ; Vc, venous cannula, with thermometer, T ; S.V.C., superior vena cava ; I.V.C., inferior vena cava, connected to the manometer M_3 for measuring the venous pressure ; R.A., right auricle ; R.V., right ventricle ; P.A., pulmonary artery ; Lu, lungs ; Tr, trachea, with cannula X^1 ; P.V., pulmonary vein ; L.A., left auricle ; L.V., left ventricle ; C, coronary artery ; Ao, aorta, ligatured ; A.C., arterial cannula in brachiocephalic artery, connected to the manometer M_1 for measuring the arterial pressure ; B, elastic cushion ; R, arterial resistance ; pressure is applied to the outside of the sleeve by the pump Pm, is stabilised by the pressure bottle P.B., and measured by the manometer M_2 ; Wa, warming coil in hot water ; X, clamp for admitting blood to the graduated vessel when it is desired to measure the output of the heart ; Tu is then temporarily clamped. (From Starling's " Principles of Human Physiology.")

valves and from the aorta by the aortic valves. The first period of ventricular systole is thus a period in which the muscular contraction is isometric (associated with no change in length). With the opening of the aortic valves the ejection phase begins and the ventricular volume rapidly diminishes. The rate of ejection gradually lessens as the ventricle empties and ceases as the aortic valves close, and the ventricle passes into diastole. With the opening of the auriculo-ventricular valves blood enters the ventricles, whose volume increases rapidly.

The Heart Sounds. The sounds associated with cardiac contraction may be heard by placing the ear on the chest wall over the heart ; they are usually detected by placing over the heart a hollow metal or vulcanite cone connected through rubber tubing to ear-pieces (the stethoscope). Each contraction of the normal heart is associated with two sounds, the first prolonged and of low pitch, the second abrupt and of higher pitch. They are usually somewhat crudely imitated by the sounds " lubb," " dup." The first sound is produced in part by muscular contraction and in part by eddies set up by closure of the auriculo-ventricular valves, the second is produced by closure of the aortic valves. Occasionally a third heart sound is heard, and is said to be due to the floating up of the auriculo-ventricular valves as the ventricle fills with blood in the early part of a long diastole.

Murmurs. When the valves of the heart are diseased, the heart sounds are usually accompanied or followed by abnormal sounds termed murmurs. Murmurs occurring with the first sound or between it and the second are associated with ventricular systole and are termed systolic. Systolic murmurs are not infrequent in normal subjects when the action of the heart is augmented, as after exercise. Murmurs occurring after the second sound and between it and the next first sound are associated with ventricular diastole and are termed diastolic ; they are purely pathological and are of first importance in detecting valvular disease of the heart. These murmurs are probably produced by the vibrations set up in the neighbourhood of diseased valves by the rapid stream of blood passing over them.

The Output of the Heart

By far the most important aspect of the physiology of the heart is the manner in which its output is adjusted to suit the demand for blood by the various organs and tissues. This may be studied most conveniently when the heart is isolated, most of the observations on the mammalian heart having been made with Starling's heart-lung preparation.

The Heart-Lung Preparation. In studying the behaviour of the isolated heart it is of great advantage to leave the lungs in full functional connection with the heart, because the blood can be aerated and the necessary oxygen supplied to the heart by ventilating the lungs, and because the lungs remove vasoconstrictor substances which develop in shed blood and make perfusion of isolated mammalian preparations difficult.

The preparation is shown diagrammatically in Fig. 1. 9, and is made briefly as follows. The venous reservoir having been filled with warm defibrinated dog's blood, a dog's chest is opened under artificial respiration ; cannulæ are tied into the brachiocephalic artery and superior vena cava and all the other systemic vessels (inferior vena cava, azygos vein, subclavian artery and

aorta) are tied. The blood entering the heart from the venous reservoir must now pass from the aorta through the arterial cannula and artificial circulation. In the artificial circulation the two important features are (*a*) the air cushion (B, Fig. 1. 9) consisting of an inverted bottle containing suitably compressed air which simulates the elastic reservoir provided by the aorta and larger arteries, and (*b*) the resistance R consisting of a thin rubber sleeve inside a glass tube containing air under the known pressure of a large reservoir with which it is connected ; blood will only flow through the sleeve at a pressure higher than that of the air outside it, and thus the arterial pressure may be kept constant and independent of output. By these two devices the blood pressure in the aorta is prevented from falling too far during diastole and so the coronary circulation to the heart is well maintained.

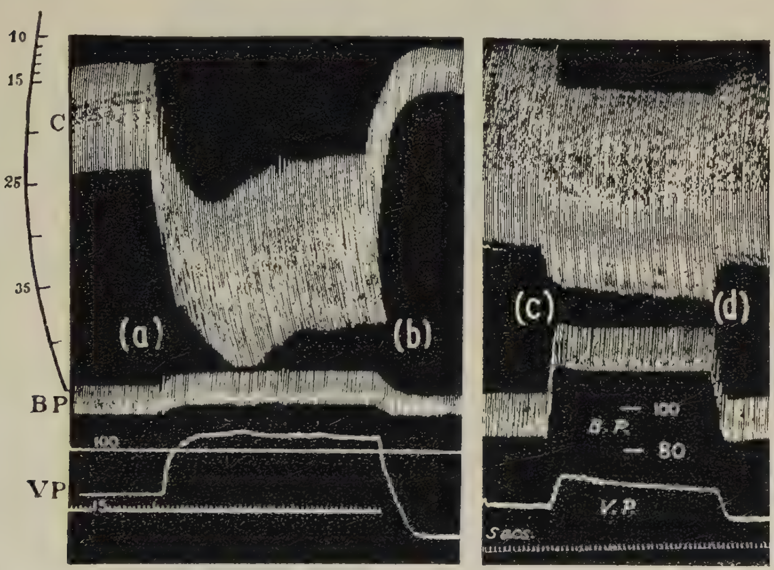
In this preparation the nerves to the heart are severed and the heart beats at a constant rate, which may, however, be varied by altering the temperature of the blood (action on the sino-auricular node).

By using the heart-lung preparation, it is possible to vary independently the venous input to the heart, and the arterial pressure developed ; the first by adjusting the height of the venous reservoir, or better by opening and closing a screw-clip on the tube between the reservoir and the vena cava, the second by adjusting the air pressure outside the rubber sleeve in the " arterial resistance." By varying the rate of venous inflow, the output of the heart can be varied smoothly over a very wide range : in this, it is totally unlike a rigid mechanical pump in which the output is fixed by the cross section of the cylinder, the throw of the crank and the number of strokes per minute. Variation of the arterial pressure, on the other hand, has no action on the output of the heart, within an upper limit set by the capability of the heart to do the necessary work, and a lower limit below which the heart is inadequately nourished through the coronary circulation. In this respect, it is quite similar to the rigid mechanical pump. The heart, in fact, acts very effectively in propelling the blood available in the venous system against any normal value of the arterial pressure ; neither allowing the veins to become engorged by propelling too little, nor sucking them empty by propelling too much.

The Mechanism by which the Power exerted by the Heart is varied. Any variation in the output of the heart (at a constant frequency) must be accompanied by a parallel variation in the amplitude of the beat—*i.e.*, in the change in length of the muscle fibres on contraction. Similarly, any variation in the arterial pressure must be accompanied by a parallel change in the force which the muscle must exert in order to expel the blood. The way in which these adjustments are made may be demonstrated by recording the volume changes of the ventricle by means of a *cardiometer*. In its simplest, and original, form this consists of a glass cup the shape of a wine glass, the stem of which is hollow and connected to a volume recorder ; the open end of the cup is fitted with a rubber membrane in which a suitably sized hole has been burned. The ventricles are slipped through this hole, the edge of which grips lightly but securely the auriculo-ventricular groove.

Fig. 1. 10 shows the results of experiments on a heart-lung preparation : a movement downward in the cardiometer tracing, C, indicates

an increase in the size of the heart; the volume at the end of diastole is thus given by the lower limit of each movement, and the volume at the end of systole by the upper limit; the total excursion gives the output of the heart per beat (stroke volume). When at (*a*) (left-hand record) the venous inflow is increased from 516 ml./min. to 840 ml./min., the venous pressure rises from 95 mm. H_2O to 145 mm. H_2O, and the ventricles distend, at first putting out less blood than they receive. The diastolic volume of the ventricles thus gradually increases, and this is associated with a gradual rise in the output per beat, until finally the output equals the inflow and the heart ceases to dilate. The arterial



(1) Venous inflow changed. (2) Arterial pressure changed.

FIG. 1. 10. Effect of changing: (1) the **Venous Inflow**, and (2) the **Arterial Pressure** on the **Volume of the Heart.**

C, cardiometer curve, the curved line at the side indicating the value of the cardiometer excursions, *i.e.* of alterations of ventricular volume, in millilitres; movement downwards indicates *increase* in volume; B.P., arterial pressure; V.P., venous pressure (water manometer) in the inferior vena cava. (Patterson, Piper and Starling.)

pressure remains practically constant, rising from 124 mm. Hg to 130 mm. Hg only, since the arterial " resistance " is specially designed for this purpose. The heart remains at the new size until at (*b*) the venous inflow is suddenly reduced to 198 ml./min.; the heart puts out more per beat than it receives, and the diastolic volume gradually falls. Finally, at a new and smaller diastolic volume the heart again puts out precisely what it receives, the venous pressure having fallen to 55 mm. H_2O.

In the right-hand record of Fig. 1. 10, the venous inflow is kept constant at 924 ml./min. When at (*c*) the arterial " resistance " is altered, so that the arterial pressure is raised from 98 to 128 mm. Hg,

the ventricles at first fail to put out as much blood as they receive and so increase in size. The increase in diastolic volume is associated with a progressive recovery in the output per beat until the heart, at an increased diastolic volume, is again ejecting all that it receives. When the arterial pressure is again reduced, at (*d*), the ventricles at first expel more than they receive and so decrease in size.

Now the output per beat—which, in a steady state must be equal to the inflow per beat—and the arterial pressure are the two factors which determine the work done by the heart on each beat : variations in either, as we have just seen, are accompanied by variations in the diastolic volume. It is found experimentally, moreover, that similar increases in the rate of work, whether produced by augmented output or by raised aortic pressure, are accompanied in a given heart by similar expansions of diastolic volume. It will be shown in Chapter 16 that the energy set free by skeletal muscle during contraction varies with the initial length of its fibres, and so it seems to be with the heart. When the diastolic volume of the heart is increased, the muscle fibres are stretched, the energy set free in systole is increased and the heart is able to perform more work. This relationship between the energy of contraction of the heart and its diastolic volume was formulated by Starling as " The Law of the Heart."

While the foregoing experiments reveal a fundamental similarity of the behaviour of heart and skeletal muscle—namely that both contract more strongly when stretched—it is now important to stress a fundamental difference. Whereas there is no mechanism in the normal body for altering the force of contraction of a skeletal muscle from a given length, such a mechanism exists in cardiac muscle, and plays an important part in modifying the action of the heart. It is mediated by the innervation of the heart (see p. 26) and by the action of the hormone adrenaline. Thus stimulation of the sympathetic nerve supply to the heart, or the administration of adrenaline, increase the force of contraction of heart muscle *from a given length.* This will be seen from the results of a heart-lung preparation experiment shown in Fig. 1. 11. Following the addition of adrenaline to the blood in the venous reservoir, the cardiometer tracing shows that the heart volume decreased. The heart performed the same amount of work from a smaller diastolic volume. It did so not only because, owing to the action of adrenaline, it beat faster and the output per beat was therefore less, but also and chiefly because the amount of energy liberated from a given length was increased.

The Venous Pressure and the Output of the Heart. It will be seen from Fig. 1. 10 that increase in inflow into the heart is accompanied by a rise in the venous pressure. This is because more pressure is required to stretch the ventricles to a greater diastolic volume. This close relationship between cardiac output and venous pressure is shown diagrammatically in Fig. 1. 12. Conversely, as shown in Fig. 1. 11, when the output is maintained and adrenaline is given, the venous pressure falls. Blood then enters the heart more readily. There

are three reasons for this: (i) smaller diastolic volume because of strengthening of the muscle and increase in heart frequency; (ii) quicker relaxation of the ventricular wall owing to the change in the properties of heart muscle caused by adrenaline; (iii) more time for the blood to enter since diastole occupies up to 10 per cent. more, and

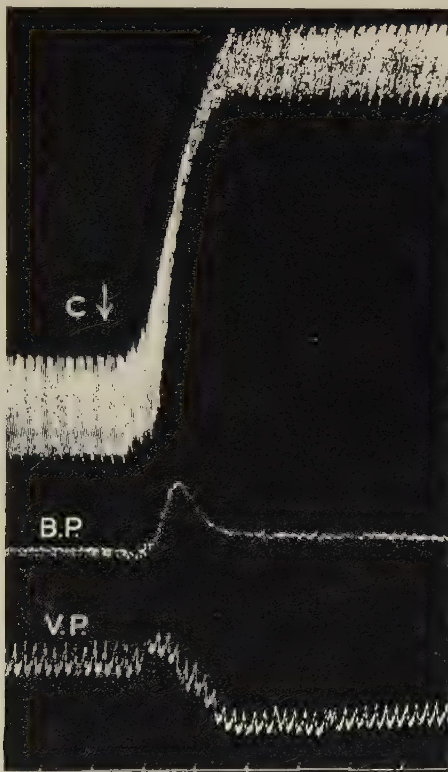


FIG. 1. 11. Effect of Adrenaline on the Volume of the Heart.

C, cardiometer curve; a rise indicates reduction in volume of the heart; V.P., pressure in the inferior vena cava.

At the arrow, 0·1 mg. adrenaline was added to the blood in the venous reservoir.

The resistance to the venous inflow was not altered. The changes in arterial blood pressure reflect the large transient increase in output while the volume of the heart was falling; and a smaller permanent increase in output owing to the increased inflow produced by the rise in level of the blood in the venous reservoir and the fall in venous pressure. (Patterson.)

systole 10 per cent. less, of the cardiac cycle. In Fig. 1. 12 have been plotted diagrammatically the relationships between cardiac output and venous pressure for a number of doses of adrenaline. It will be seen from such a "family of curves" (after Sarnoff): (a) that increase in venous pressure increases cardiac output; and (b) that for any given venous pressure the cardiac output is greater during the action of adrenaline.

If the venous inflow is increased and at the same time the heart beat is quickened and strengthened, then an increase in cardiac output could occur with little or no change in venous pressure or diastolic volume. This explains why the increase in cardiac output during exercise may not be accompanied by any change in venous pressure or in heart size (see later, pp. 49, 52).

As previously stated, the inflow into the heart in the conventional heart-lung preparation is determined almost wholly by the resistance to the passage of blood past the screw-clip placed on the tubing between

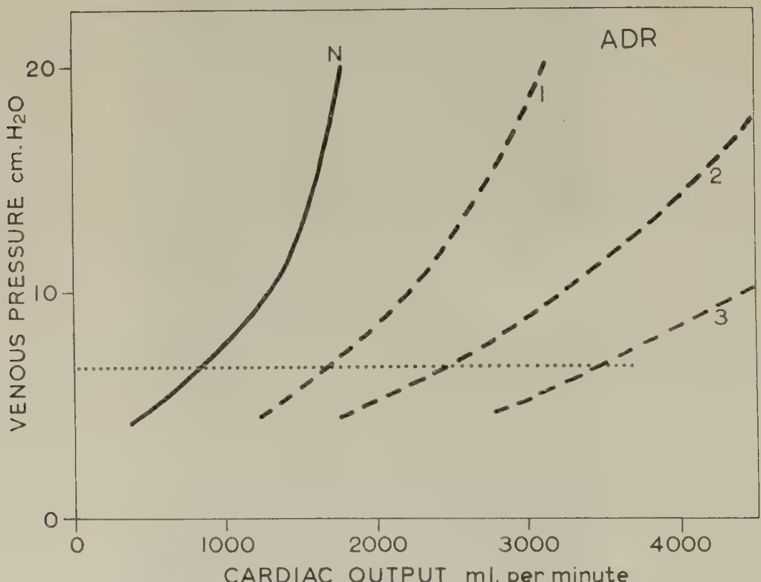


FIG. 1. 12. Diagram showing the Relations between Venous Pressure and Cardiac Output.

Observations on a heart-lung preparation.

N, normal condition. Increase in venous pressure causes increase in cardiac output. For a venous pressure of 8 cm. H_2O, the output is approximately 800 ml./min.

ADR 1, 2, 3, after addition of successively greater amounts of adrenaline, causing quickening and strengthening of the beat. For a venous pressure of 8 cm. H_2O, the cardiac output became 1,300, 2,500 and 3,500 ml./min. respectively.

the venous reservoir and the heart (Fig. 1. 9, Cl), this resistance being large compared with that opposed to the entry of blood into the heart itself. Accordingly, changes in frequency and strength of the beat can have little effect on the inflow and thus on the output of the heart. An altogether different situation arises, however, if the screw-clip is removed and the venous reservoir quickly lowered until the level of the blood in the reservoir is the same as that in the venous manometer. In these circumstances, owing to the capacity of the reservoir, the level of blood and the venous pressure must remain nearly constant. The

main resistance to the flow of blood into the heart is now that of the heart itself. In this preparation adrenaline causes a large increase in output, and this is not surprising as the heart fills much more readily for the reasons that have already been stated.

The Work Done by the Heart. The external work done by each ventricle during a single contraction is the sum of the potential and kinetic energy imparted to the expelled blood. In symbols, work done per minute $= (P.V. + \frac{1}{2} \rho Vv^2) N$, where P represents the mean aortic pressure, V the stroke volume or output per beat, v the velocity of blood in the aorta,[1] ρ the density of the blood and N the number of beats per minute.

Assuming that the pressure developed in the right ventricle is one-sixth that of the left, the external work done by the whole heart per minute becomes : $V (\frac{7}{6} P + \rho v^2) N$.

The oxygen consumption of the heart beating at constant rate is found to be proportional to the external work done and thus to the diastolic volume of the heart. If the heart rate is varied, then it is found that the oxygen consumed per unit of work done is greater at a high rate of beat than at a low. Since the oxygen consumption is a measure of the total energy liberated, we may say that the efficiency of the heart, $\frac{\text{external work done}}{\text{energy set free}}$, is greater at low rates than at high. The efficiency of the isolated heart is usually about 0·2 (20 per cent.).

The Output of the Heart in Man

In determining the circulation rate in man, indirect methods alone are available, and unless these methods are to defeat their ends it is essential that they should not of themselves alter the cardiac output. What is, in fact, actually measured is the volume of blood which flows through the lungs in a given time. The general principle of the methods used is quite simple, and similar methods are used to measure the rate of blood flow in any organ, as we shall see later. In passing from the right side of the heart to the left, some substance X is added to, or removed from, the blood at a measured rate of, say, Q grams, or millilitres, per minute. The actual concentration of X in the pulmonary arterial and venous bloods are measured and found to be c_a and c_v g./l. (or ml./l.) respectively. Each litre of blood thus gains or loses $(c_a - c_v)$ grams (or millilitres) of X, so that if CO is the cardiac output in litres per minute, $CO (c_a - c_v) = Q$, or $CO = Q/(c_a - c_v)$.

This general relation is known as the **Fick Principle.** In the lungs, it is applied to the gain of oxygen or the loss of carbon dioxide ; Q being now the volume of oxygen absorbed per minute (or of carbon dioxide evolved), in litres per minute, c_a being the volume of oxygen

[1] This is derived by simple calculation from the stroke volume, the area of the opening through the aortic valves, and the ejection time (usually about three-eighths of the duration of a complete cardiac cycle). At rest, the velocity is about 50 cm./sec.

(or of carbon dioxide), in millilitres, contained in each litre of arterial blood, and c_v the volume contained in each litre of venous blood.

For example, in a normal subject engaged in sedentary occupation such as reading, the volume of air taken into and out of the lungs per minute (the " respiratory minute volume ") is about 8 litres, or 8,000 millilitres. Since inspired air contains 21 per cent. oxygen and practically no carbon dioxide, while expired air contains 16·4 per cent. oxygen and 4·1 per cent. carbon dioxide (Table 4. 1, p. 106, all gases being analysed dry), the amount of oxygen absorbed is approximately 8,000 $(21 - 16{\cdot}4)/100 = 370$ ml./min., and the amount of carbon dioxide liberated is approximately 8,000 $(4{\cdot}1 - 0)/100 = 330$ ml./min. In Figs. 3. 1 and 3. 3 (pp. 77 and 88) the oxygen and carbon dioxide contents of the arterial blood are shown as 19·5 ml./100 ml., and 50 ml./100 ml., respectively, while for the venous blood of a " resting " subject, the corresponding values are 11·7 ml./100 ml., and 57 ml./100 ml., respectively. These figures must all be multiplied by 10 so as to keep the units consistent throughout. The circulation rate may now be calculated from the Fick equation. (1) From the oxygen figures, the circulation rate is $370/(195 - 117) = 4{\cdot}8$ litres per minute. (2) From the carbon dioxide figures, the circulation rate is $330/(570 - 500) = 4{\cdot}7$ litres per minute.

These two values of the circulation rate do not agree exactly, as they should do, since they are derived from representative, and not highly accurate, data. The values derived from the uptake of oxygen are in any case slightly too small, since no allowance has been made for the fact that the volume of carbon dioxide liberated is less than the volume of oxygen absorbed, the R.Q. being ordinarily less than 1 (Chapter 6) : the volume of air inspired, therefore, must be slightly greater than the measured volume of air expired. (The method of calculation is described in Chapter 6, p. 182.) For this allowance to be of importance, one must measure accurately all the desired quantities in the particular subject at the time of investigation.

In practice this method is used as follows. Since the composition of the inspired air is known, the volume of oxygen absorbed and of carbon dioxide liberated may be obtained by collecting the expired air over a known period, measuring it and estimating its oxygen and carbon dioxide content. Arterial blood may be obtained from puncture of an artery and analysed for its oxygen and carbon dioxide content.

If the oxygen or carbon dioxide pressure in a specimen of alveolar air is determined, the oxygen or carbon dioxide content of the arterial blood may be calculated from the dissociation curve of blood, after the total oxygen capacity has been determined on a specimen of venous blood (see Chapter 3). This method is not valid in diseased subjects, since in certain conditions it is untrue that the arterial blood is in gaseous equilibrium with the mixed alveolar air.

The real difficulty of this method lies in determining the gaseous content of the venous blood, for the composition of venous blood from different parts of the body is variable, and the venous blood here in question is the mixed venous blood entering the lungs. In animals the right ventricle may be punctured and venous blood withdrawn for

analysis, but to do this in man is unjustifiable. However, blood from the right auricle and ventricle can be obtained in man by means of *cardiac catheterisation.* A radio-opaque unwettable cardiac catheter is introduced through a wide bore needle into the median anti-cubital vein of the left arm and passed up through the innominate and superior caval veins into the right auricle where its position is verified by X-ray. It may be left *in situ* for an hour or more. A slow infusion of 3·8 per cent. sodium citrate solution through the catheter prevents clotting round the hole at its tip. Samples of right auricular blood are withdrawn into a syringe under oil, and their oxygen content estimated. The oxygen consumed by the subject per minute is also estimated.

An older and indirect method of estimating the oxygen and carbon dioxide contents of the mixed venous blood depended on the following reasoning. It is clear that if we stop breathing, the gas in the lungs will gradually tend to come into equilibrium with the venous blood, the oxygen content falling and the carbon dioxide content rising. Unfortunately we cannot measure the composition of the venous blood as simply as this would suggest, because equilibrium takes too long to be established. If this method is to be used at all, the whole operation must be completed in a time less than that taken for any part of the blood to circulate once (about 23 seconds at rest) ; if a longer time is taken, then the composition of the venous blood will be altered, because such blood before traversing the tissues and returning to the lungs will already have been equilibrated with an abnormal gas mixture in the lungs. The difficulty can be overcome in several different ways, the use of **intermittent rebreathing** being one of the best. In principle, the subject breathes in and out of a bag containing air, for about 10 seconds or so : the bag is then closed and the subject breathes from the open air for a while, so as to get the composition of the blood back to normal again. He then rebreathes from the bag for a further short period, and so on. During each rebreathing, the composition of the air in the bag approaches that of the mixed venous blood, and eventually reaches it. In practice, the bag is made to contain about 6 per cent. carbon dioxide initially, so as to reduce the time required for equilibration.

The Foreign Gas Principle. The necessity for estimating the concentration of the reference substance X in the venous blood is avoided altogether if this substance is foreign to the body, and is removed completely as the blood passes through the tissues. No such substance has yet been discovered, however (except in the special case of the kidney, as will be seen later). But if the whole sequence of procedures can be carried out in less than the circulation time—*i.e.*, before any significant quantity of X has returned to the lungs in the venous blood—the same end is achieved. Although nitrous oxide and ethylene have been used as the foreign gas, acetylene appears to be the most suitable for this method, since it is harmless and easily estimated, diffuses readily through the lungs, and has a convenient and constant solubility in the blood. After emptying the lungs, the subject breathes quickly and deeply four times in and out of a rubber bag containing 2 litres of air and 0·5 litres of pure acetylene, a sample of the air being taken at the end of the last expiration. The lungs are thus filled with a gas mixture containing a known percentage of acetylene. Another sample of the gas is taken after two to six more breaths in and out of the bag, during which time some of the acetylene is carried away in the arterial blood. From the composition of the two samples, and from the total quantity of gases in the lungs and bag, the amount of acetylene absorbed and its average partial pressure in the alveolar air (and thus its concentration in the arterial blood) are determined.

The Hamilton Dye Injection Method. A known amount of the dye " cardiogreen " is rapidly injected into a vein. During its passage through

the pulmonary circulation, it becomes evenly distributed throughout the blood stream. Successive samples of arterial blood are collected, by means of a disc rotating at a known speed, and having attached to its circumference a large number of small tubes which in turn catch the blood flowing from an intra-arterial needle. Suppose, as an ideal simplification, that the dye first appears in the arterial blood at a time t_1 seconds after the intravenous injection, and disappears again at a time t_2 seconds. Then, if CO is the cardiac output in litres per second, the total volume of the blood in which the dye becomes distributed is $CO \times (t_1 - t_2)$. If the arterial concentration of the dye in the interval between t_1 and t_2 is c_a, and D grams of dye were injected, then $D/CO(t_1 - t_2) = c_a$ or $CO = D.60/c_a(t_1 - t_2)$, in litres per minute. Actually, the calculation is more complicated than this, since the dye does not appear in the arterial blood and disappear again, suddenly, nor is its concentration constant. The successive estimations of c_a must be plotted against time, and the mean value of the arterial concentration c_a and the best values of t_1 and t_2 determined from the curve. This method has the advantage that the subject is not required to co-operate with the experimenter in carrying out special respiratory procedures.

The output of the heart is remarkably constant for each individual under conditions of complete physical and mental rest, that is after the subject has been lying down and fasting for ten or more hours (basal conditions). In different individuals the output is closely related to the surface area of the body ; and varies from 4 to 5 litres per minute (2·5 l./min. for each square metre of body surface), as estimated by the acetylene and dye injection methods, though rather higher values have been obtained by the method of cardiac catheterisation.

The cardiac output is reduced by a change from the recumbent to the standing position, which produces a fall in right auricular pressure. It is unaltered during sleep. Small increases are produced by excitement (about 1 litre per minute), and by the ingestion of food and drink (up to 2 litres per minute) ; but it is in muscular exercise that the greatest increases are found. Even such slight exercise as flexing the thigh once every second doubles the cardiac output. With more severe exercise, such as cycling, running, or swimming, the cardiac output increases proportionately to the rate of work, commonly reaching 20 litres per minute, and in trained individuals values of over 30 litres per minute have been attained. This will be discussed further in the next chapter (pp. 48–52).

The Innervation of the Heart. The heart receives two sets of nerve fibres from the autonomic system, parasympathetic fibres from the vagus and sympathetic fibres (see Chapter 15). The vagal fibres terminate in ganglion cells in the heart, from which fibres pass to the sino-auricular and auriculo-ventricular nodes and the auricular muscle. Stimulation of the vagus slows the heart by action on the pace-maker, and also depresses the conduction from the auricle to the ventricle by action on the auriculo-ventricular node ; it may also reduce the force of the auricular contractions, but in the mammal has little action on the ventricles themselves. The sympathetic fibres arise from cell stations in the middle and inferior cervical ganglia and terminate around the sino-auricular and auriculo-ventricular nodes and in the heart muscle. Stimulation of the sympathetic fibres quickens the heart by action on the pace-maker,

facilitates conduction from auricle to ventricle, and augments the force of the auricular and ventricular contractions. It may thus be seen that the action of the vagus on the heart is almost the converse of that of the sympathetic (reciprocal innervation). Normally, impulses are passing to the heart along each set of nerves, for section of the vagus quickens the heart and section of the sympathetic slows it.

The Control of the Frequency of the Heart Beat. The rhythm of the pace-maker is affected by impulses in the vagal and sympathetic nerves, by the concentration in the blood of adrenaline and thyroxine Chapter 11), and to some extent by the temperature of the blood. The nervous impulses come from the cardiac parts of the vaso-motor centre (Chapter 2, p. 42). The activity of this centre depends upon local conditions and upon reflexes from the cardio-pulmonary system and from receptors (*baroreceptors*) in the walls of the aorta and carotid artery which respond to a rise in arterial pressure (see Chapter 2, p. 42). In the normal resting subject, the natural rhythm of the pace-maker is made slower by vagal impulses from the centre : if the vagal nerve endings are paralysed by atropine (belladonna), the heart rate increases from about 70 to about 180 beats per minute. This is due largely to the fact that the baroreceptors are normally in a state of moderate excitation, one of their actions being to bring about a reflex slowing of the heart (compare Fig. 2. 8, p. 47).

The heart rate is *increased* in the following circumstances:

(*a*) In a warm environment, and by fevers.

(*b*) By hæmorrhage, and when a recumbent subject stands upright.

There is a reduction in venous return, owing to an actual reduction in blood volume, or to blood pooling in the lower part of the body ; the cardiac output decreases and the arterial pressure tends to fall. The baroreceptors—and thus the cardiac centre—are less excited, there is a reduction in vagal inhibition and an increase in heart rate.

(*c*) By excitement, fright, exercise and lack of oxygen. This is due, at least initially, to an action of the higher centres of the brain on the vasomotor centres, reducing or abolishing the normal vagal inhibition of the heart. Later, in exercise, reflexes from the cardio-pulmonary system, the increase in the amount of circulating adrenaline, and the increase in the temperature of the blood accentuate and prolong the increase in heart rate. Reflex action through the vasomotor centres contributes, also, to the effects produced by breathing air deficient in oxygen ; but the site of the receptors has not yet been discovered. The arterial pressure rises both in exercise and during lack of oxygen : reflexes from the baroreceptors will thus tend to slow the heart, but their effect is swamped by the predominant accelerating action of the other factors.

The heart rate is *decreased* greatly during a fainting attack, and in some subjects when consciousness is lost as a result of severe lack of oxygen (p. 53).

Increase in the intracranial pressure, as. for example, by a cerebral tumour,

often slows the heart; and this is believed to be due to restriction of cerebral flow and to the effect of lack of oxygen on the vagal centre. Adrenaline injections in pregnant rabbits slow the fœtal heart; this is because constriction of the uterine vessels causes oxygen lack in the fœtus; the mechanism for accelerating the heart does not function, and so the effect of the oxygen lack on the pace-maker becomes manifest.

Heart Failure. The function of the heart is to propel blood through the vessels at a rate fast enough to meet the metabolic requirements of the tissues; failure to do so may result either from a lesion of the heart itself or because the supply of blood to the heart from the great veins is inadequate (peripheral circulatory failure). Heart failure in man is frequently accompanied by a condition of stenosis (narrowing) or incompetence (leakiness) of one or more valves, the heart then working at a mechanical disadvantage; yet it will readily be appreciated that the origin of heart failure is usually to be sought in disease of the heart muscle itself. The earliest symptom of heart failure is undue breathlessness on exertion. In advanced heart failure the patient is breathless at rest and more so in the recumbent than the sitting posture; venous pressure in the cervical veins is increased, the liver is enlarged and extensive accumulations of fluid occur, in the tissues (œdema) of the legs and in the peritoneal (ascites) and pleural cavities. In such an advanced case the cardiac output may be reduced to half the normal. This reduction is not the chief cause of the symptoms since a similar or greater reduction of cardiac output in peripheral circulatory failure is associated with an entirely different clinical picture. The chief factor in producing the symptoms in cardiac failure is the rise in venous pressure which, on the left side of the heart, produces engorgement of pulmonary veins and capillaries, and by thus increasing the rigidity of the lungs leads to breathlessness and may lead to œdema of the lungs; on the right side of the heart the raised venous pressure distends the hepatic veins and thus enlarges the liver, and by raising capillary pressure favours the passage of fluid out of the capillaries into the tissues.

The most instructive example of cardiac failure in man is that seen in paroxysmal tachycardia in which the heart suddenly begins to beat at a rate of 180 to 200 beats per minute. As we have seen (p. 23 above), for a given output and arterial pressure, the rate of oxygen consumption by the heart rises as the frequency of the beat becomes greater. The flow of blood in the coronary vessels, moreover, occurs chiefly during diastole: if the duration of diastole is reduced excessively, the rate of supply of oxygen will fall, and may well become inadequate. At very high frequencies, then, even previously healthy hearts may after some hours display the phenomena of cardiac failure. With the onset of the paroxysm the patient gradually becomes breathless, the neck veins swell, the liver enlarges, and the cardiac impulse moves outwards for perhaps 2 inches. This movement of the impulse is due to the enlargement of the diastolic size (cardiac dilatation) consequent on the increased venous pressure. Even with such a cardiac dilatation the output of the heart is found to be reduced. With the end of the paroxysm, the neck veins collapse, the liver decreases in size and the heart's impulse returns to its normal position; the relationship between cardiac output, diastolic size and venous pressure once more becomes normal.

The Arterial Pulse

The sudden ejection of blood into the aorta that occurs with each beat of the ventricles produces a wave of increased pressure that is propagated along the arteries towards the periphery; this is known as the pulse wave, and it may be felt and recorded in any of the superficial arteries of the body as the pulse beat. The velocity at which this pressure wave is propagated may be determined by recording the

times of its arrival at two different points such as the subclavian and radial arteries, and dividing the time differences by the distance between the two points at which measurements are taken. The pulse wave velocity varies in different subjects chiefly with the thickness and elasticity of the arterial wall, and since the arteries tend to become more rigid with advancing years (arterio-sclerosis), the velocity increases from an average rate of 5·2 m. per second at the age of five to an average of 8·6 m. per second at the age of eighty-four. It is important not to confuse the pulse wave, which is simply a wave of increased pressure, with the movement of the blood itself ; the velocity of blood-flow is nowhere greater than 0·5 m. per second at rest and is considerably less in the smaller vessels.

The Form of the Pulse. The changes in pressure and their spacing in time that occur in a superficial artery such as the radial may be recorded by the *sphygmograph*, an instrument in which the pulsations of the artery are transmitted to a lever writing on a piece of smoked paper moving at a suitable speed. With practice the main features exhibited by such a record can usually be ascertained with the finger. The main deflection in the pulse record, the primary wave (Fig. 1. 5), is the result of the sudden distension of the aorta during ventricular systole. Following the primary wave are a number of secondary waves, which arise in several ways. The most constant and conspicuous secondary wave results from closure of the aortic (semi-lunar) valves. At the end of ventricular systole the pressure in the ventricle falls rapidly, and blood begins to flow back from the aorta, but is suddenly checked by closure of the aortic valves. This sudden check causes a rise of pressure in the aorta, and this wave travels down the arterial tree with the same velocity as the primary wave which it follows. This secondary wave is known as the dicrotic wave and is preceded by a notch, the dicrotic notch. From what has been said it will be realised that the upstrokes of the primary and of the dicrotic waves are separated by the same interval of time as the opening and closure of the aortic valves. In addition to those two waves there are a number of inconstant and small waves arising from the reflection of the primary wave by obstacles such as the bifurcations of the arteries.

The form of the pulse wave is modified by such conditions as affect the discharge of the blood from the heart and its escape through the arteries. Thus, when the aortic valves are narrowed by disease (aortic stenosis) the distension of the aorta during systole is very slow and the pressure in the peripheral arteries rises slowly in the maximum ; this slow-rising pulse is termed anacrotic. When, on the other hand, the aortic valves do not close properly (aortic regurgitation), the pressure in the aorta falls very quickly during diastole and is suddenly and greatly raised during systole ; in this condition the upstroke of the primary wave is unusually sudden and its downstroke rapid ; from the sudden thrust on the finger feeling the artery this pulse is described as "water hammer." In conditions of low pressure and rapid blood-flow, as may occur in children and in fevers, the dicrotic wave may be so pronounced as to be easily felt at the wrist ; the frequency of the heart may thus be mistaken for twice its true value.

CHAPTER 2

THE CIRCULATION—CONTROL OF BLOOD FLOW

On each contraction of the heart, blood is expelled from the ventricles in jets at high pressure into the aorta and pulmonary artery. The **aorta** is a wide tube, the thick walls of which are largely composed of elastic tissue ; like any other elastic structure, its capacity is determined largely by the pressure of the blood it contains. When blood is expelled during ventricular systole, the aortic pressure rises, the aorta is distended and so accommodates a large part of the blood expelled, the remainder escaping through the arteries. During ventricular diastole, the tension in the aortic walls maintains the flow of blood onwards through the arteries, and the aorta diminishes in size until it is again distended at the next heart beat. In this purely passive way the aorta (and to a less extent its larger branches, which are similar in structure) converts the intermittent flow from the heart into a continuous though pulsating flow in the arteries.

The **arteries** have smooth muscles in their walls, which, on suitable stimulation, will contract actively or relax. The proportion of muscular to elastic tissue is greater in small arteries than in large, and the diameter, particularly of the smallest branches, or **arterioles**, can change over a wide range independently of the pressure within the lumen. The state of constriction or dilatation of the arterioles supplying any particular organ largely determines the proportion of the total cardiac output which is sent to it.

The **capillaries** are about 10μ in diameter and the blood they contain is separated from the tissues by a single layer of flat endothelial cells which forms the capillary wall ; it is accordingly here that the interchange of substances between the blood and the tissues takes place. In spite of their thin walls and the absence of muscle cells, the capillaries are capable of active contraction, and of exerting pressure of 60 mm. Hg or more when contracted. In some resting tissues the majority of the capillaries are closed ; during activity they open, and thus they also play a part in regulating the distribution of the blood to the organs. Although the capillaries are under nervous control, they are pre-eminently the vessels which react to chemical substances released during the activity of tissues which they supply.

The blood from the capillaries is collected into **venules**, which join up to form **veins.** These are wide and relatively thin-walled, and offer little resistance to the flow of blood. They are capable of active variation of calibre and are under nervous control. All but the smallest and largest veins contain valves, consisting of a number of semi-circular folds of the intima projecting into the lumen. As a rule two such folds are placed opposite one another, and are so formed that when the blood is forced in a direction away from the heart, the folds float out into the blood-

stream and block the vein. When the muscles contract the thin-walled veins are squeezed and blood is forced in the only direction it is free to travel, namely towards the heart. When the muscles relax, blood can enter the veins, but only from the arterial side. This " muscle pump " is an important mechanism for facilitating the venous return to the heart, as will be discussed later (p. 51).

The artery which supplies an organ, and the vein which drains it, may be some 2 to 5 mm. in diameter ; there will be 10,000 to 100,000 arterioles and venules, 0·02 to 0·05 mm. in diameter ; and millions, or tens of millions, of capillaries, some 0·01 mm. in diameter. As the vascular tree branches, the total cross sectional area of the vessels increases : that of all the arterioles is around 10 times that of the artery, and that of all the capillaries some 10 times greater again. Thus the velocity of blood, about 10 centimetres per second in the artery, decreases progressively until, in the capillaries, it is less than 1 millimetre per second ; it then rises again and reaches several centimetres per second in the vein. Even though each capillary is very short (less than 1 mm. long) the average particle of blood spends enough time in one (about 1 second, or less if the organ is active) for interchange to take place with the tissues. (All these figures are very approximate, for purposes of illustration only.)

The Flow of Liquids in Tubes. Viscosity

In 1836 and the following years, Poiseuille performed experiments on animals with a view to discovering the relations between blood pressure and flow in the circulation. He found his observations so bewilderingly variable that he turned his attention to the simpler problems concerned with the flow of pure liquids like water through glass tubes. For our purposes, the results of his experiments—which are embodied in Poiseuille's Law—may be summarised by the following statements.

(*a*) The rate of flow of a liquid through a tube is proportional to the pressure driving it : the ratio of the pressure to the rate of flow is the *resistance* of the tube to the flow of the liquid used. The flow of a liquid through a tube, therefore, is exactly analogous to the flow of an electric current through a wire, as expressed in Ohm's law, $i = E/R$. This statement ceases to be true if the rate of flow is very large and the flow becomes " turbulent." Local turbulence may occur in the blood vascular system, but it is not sufficient to affect the pressure–flow relations.

If we have a number of tubes with different resistances arranged in such a way that all the fluid goes through all of them in succession (in " series "), the pressure drop across each will be proportional to its resistance, and the total resistance of the whole will be the sum of the separate resistances. If the tubes are arranged in such a way that the flow is divided among them (in " parallel "), and all have the same pressure across them, the flow through each will be inversely proportional to its resistance, and the reciprocal of the total resistance of the whole—*i.e.* the total *conductance*—will be the sum of the separate

conductances. Any complicated system of tubes in series and in parallel (such as the blood vascular system) will thus have a resistance to flow (the " total peripheral resistance ") which will depend on the separate resistances of its component parts ; but it may always be measured in terms of the ratio of the pressure applied to the total rate of flow produced.

(*b*) The resistance to flow of any particular liquid is directly proportional to the length of the tube (l), and inversely proportional to the fourth power of its radius (r) : the resistance thus increases very rapidly as the radius becomes smaller. This accounts for the fact that the resistance of the arterioles is much greater than that of the arteries (see p. 39), in spite of the fact that the total cross sectional area is greater and the velocity of flow smaller.

(*c*) For a given tube, the resistance to flow depends on the nature of the liquid driven through it, and the ratio : (resistance to flow of liquid X) divided by (resistance to flow of water) is a characteristic property of liquid X known as its *relative viscosity* (r) (*i.e.* viscosity relative to that of water). The relative viscosity is the same even if the length or diameter of the tube, or the applied pressure and rate of flow, are varied over wide limits.

Expressed in symbols, we may thus write Poiseuille's law :

$$P/F = R = 8\eta l/\pi r^4$$

Poiseuille's law is not always exactly obeyed by blood, since, unlike that of " perfect " liquids, the value of its viscosity may depend on the rate of flow and on the dimensions of the tube. Strictly, therefore, the idea of viscosity as a " constant " property of a liquid is inapplicable to blood, and the value obtained under any particular set of conditions is called the *apparent viscosity*.

(i) If blood is made to flow through a tube at a sufficiently high velocity, by applying a sufficiently high pressure, its apparent viscosity will be the same whatever actual velocity or pressure is used : it behaves as a " perfect " or " Newtonian " fluid. But if the velocity and pressure are reduced considerably, this is no longer true : the apparent viscosity becomes progressively larger as the velocity and pressure are made smaller, and the flow is now " anomalous " or " non-Newtonian." It is probable that if the rate of flow is small, the red blood cells stick together and an extra force is required to tear them apart and to allow the blood to flow ; the resistance to flow, and the apparent viscosity, are thus increased. If the rate of flow is very large, they move past each other so fast that they cannot stick together, this extra force is no longer needed, and the resistance to flow, and apparent viscosity, become smaller.

In the blood vessels of a normal animal or man, the velocity is sufficiently great for the apparent viscosity to be independent of the rate of flow, and we are justified in regarding the resistance of the blood vessels as a quantity which varies only with their calibre. Clinical measurements of the viscosity of blood, however, are sometimes made in an " Ostwald viscometer "; the movement of the blood is then so slow that the apparent viscosity is largely affected by the rate at which the blood flows through it—a matter depending on the exact design of the instrument. In a " Hess viscometer ", the blood is made to flow at a velocity so great that small variations do not affect the apparent viscosity.

(ii) If the diameter of the tube through which the blood is made to flow is greater than about 0·2 mm., the apparent viscosity (in any conditions of flow) is independent of the actual value of the diameter. But if the diameter is less than about 0·2 mm., the apparent viscosity becomes smaller as the diameter is reduced still further. It is important to remember that the *resistance* to flow becomes greater as the diameter becomes smaller, whether the fluid used is " perfect " or has this anomalous property shown by blood : but the increase in resistance, for a given decrease in diameter, is smaller with blood than with a " perfect " fluid. This effect is due to the presence of a narrow layer in contact with the wall of the tube which is deficient in red cells and so has a smaller viscosity than the rest of the blood in the tube. The width of this layer (about the thickness of a red cell) is practically the same whatever the diameter of the tube. In relatively large tubes, its effect is negligible : but in small tubes, it provides a " lubricating " layer which becomes progressively more effective in allowing the blood to slip through the tube, as the diameter of the tube becomes progressively smaller.

It is with very small tubes, namely the arterioles, that we are mainly concerned as physiologists ; as will be seen in Fig. 2. 4 (p. 40 below), the main pressure fall between the arteries and veins occurs in the small blood vessels. In tubes of such a size (about 0·02 mm. diameter) the apparent viscosity has only about one-half the value found in large tubes. Halving the apparent viscosity of the blood would be an important economy in the circulation, for it should halve the work done by the heart in maintaining a given circulation rate ; for a given maximum output of work of the heart, it would about double the amount of work a man could perform in violent exercise. Moreover, in the circulation the resistance of the blood vessels is varied by changing their diameters. The fall in apparent viscosity with reduction in diameter allows this control of resistance to be more delicate than it would be if Poiseuille's law were accurately obeyed.

Representative values of the apparent viscosity of blood are given in Table 2. 1. They include about the smallest and the largest values which have been obtained on a given sample of blood, and show how greatly the viscosity may vary according to the conditions of measurement. In any conditions of flow and size of tube, the viscosity of blood increases rapidly with increase in the hæmatocrit value, particularly

TABLE 2. 1

The Apparent Viscosity of Blood (relative to water)
(Representative Values)

Conditions of Measurement	Hæmatocrit Value (Per cent.)			Remarks
	30 (anæmia)	45 (normal)	60 (high-altitude acclimatisation)	
Rate of flow sufficiently large				
In ordinary tubes (diam. greater than 0·2 mm.)	3·0	4·5	6·5	Independent of exact conditions of measurement.
In very small tubes, *e.g.* arterioles	2·0	2·2	3·1	Depends on size of tube.
Rate of flow very small (large tubes)	15	40	130	Falls rapidly with increase in rate of flow.

when this exceeds about 50 per cent. ; the figures given in Table 2. 1 show, roughly, the range of physiological interest. This is the most important factor affecting the viscosity of the blood in the circulation.

The *absolute* viscosity of blood, and thus the resistance to flow through a given tube, rises as the temperature falls. The effect is not large, and in ordinary conditions of life, the variation never exceeds a few per cent. At the lowest temperature compatible with life (about 23° C.), the viscosity is about 40 per cent. greater than in normal conditions : and at the highest temperature (about 44° C.), it is about 15 per cent. smaller. The effect of temperature on the viscosity of blood is nearly the same as that on the viscosity of water, so that the *relative* viscosity of blood is sensibly independent of temperature.

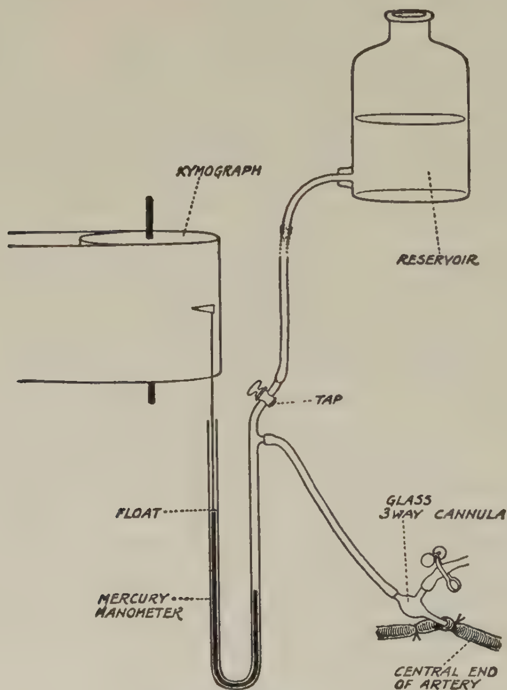


FIG. 2. 1. **Apparatus for recording Blood Pressure in Animals.**

The Blood Pressure

If the effects of respiration and of muscular movement on the veins are excluded, the flow of blood through the vessels is produced entirely by the pressure differences established by the heart. Although the blood pressure thus falls continuously from its highest value in the aorta to its lowest in the great veins entering the heart, it is most conveniently measured in three situations, the larger arteries, the capillaries and the great veins.

The Arterial Blood Pressure. The pressure in the arteries varies with each heart beat, and thus has a **maximum or systolic** value and a

minimum or diastolic value ; the difference between the two is termed the **pulse pressure.** In the experimental animal the arterial pressure is commonly measured by inserting into the carotid or femoral artery a glass cannula filled with a solution which prevents the blood from clotting (*e.g.* half-saturated Na_2SO_4), and connected through pressure tubing filled with the same solution to a mercury manometer (Fig. 2. 1). Since the flow is obstructed in the vessel cannulated, the pressure recorded is that at the point of junction with the larger vessel that supplies it ; for since there is no flow along the obstructed vessel there is also no fall of pressure. The levels to which the mercury rises in systole and diastole do not accurately represent systolic and diastolic pressures, since they are largely determined by the momentum of the heavy column of mercury. When true systolic and diastolic pressures are required the arterial cannula must be connected through a liquid system to a manometer that will respond quickly enough to record the pressure changes without lag, and in which the actual movement of the recording system is very small, so that a negligible volume of blood is withdrawn from the circulation. A stiff membrane on the end of a tube is ordinarily used, and its deflection is magnified either optically, a small mirror being mounted on the membrane, or electrically, a " pick-up " sensitive to changes in electrical resistance (" strain-gauge manometer ") or capacitance (" capacitance manometer ") being used. The magnification obtained by this method is so large that the volume of blood displaced in the tubing by an extreme pressure change is less than the volume of a single red blood corpuscle. When the arterial cannula is connected to a mercury manometer of very wide bore and hence of very low frequency, the mercury column does not oscillate and records the **mean pressure** of the blood ; more usually this value is recorded by using a mercury manometer of narrow bore with very high damping. The mean arterial pressure can be calculated from the systolic and diastolic values if the form of the pulse wave is also known ; it approximates more closely to the diastolic than to the systolic pressure, since diastole lasts longer than systole.

For clinical investigations, the arterial pressure in man may be measured by similar methods. Under local anæsthesia a needle is inserted into the brachial or femoral artery and connected to a capacitance manometer. For routine work an indirect method based on the following principle is used. If an artery is compressed, then the minimum pressure serving completely to stop the flow must be at least as great as the highest pressure (systolic) attained inside the vessel. As the pressure outside the artery is reduced, so blood will flow through the artery for longer and longer periods of the cardiac cycle, until finally when the compressing force is just less than diastolic pressure, blood-flow will be unimpeded, and the artery will cease to be deformed at any point of the cardiac cycle. A flat rubber bag contained in a loose but inextensible silk case is wrapped snugly around the upper arm. The interior of the bag is connected to a small hand pump through which air can be introduced or removed, and to a mercury manometer

which measures the pressure of the air in the cuff. The cuff must be sufficiently wide to transmit the pressure of air it contains to the centre of the limb (12 cm. for the upper arm) and sufficiently long to encircle the limb completely. The systolic and diastolic pressures are determined with this **sphygmomanometer** (Fig. 2. 2) as follows.

(1) **By Palpation.** The cuff is inflated until the pulse can no longer be felt at the wrist. Air is allowed to leak out until the pulse returns. The pressure at which the pulse beat can first be felt is taken as systolic pressure.

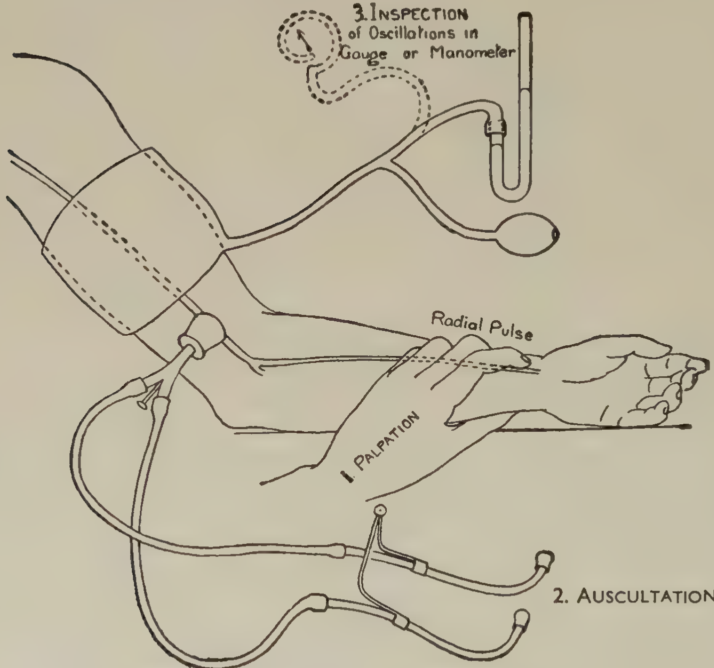


Fig. 2. 2. **The Measurement of the Arterial Blood Pressure in Man.** (From Harris' "Experimental Physiology.")

(2) **By Auscultation.** The bell of the stethoscope is placed over the brachial artery at the bend of the elbow, and the cuff on the upper arm inflated until all sounds disappear. Air is allowed to leak out until pulse sounds just reappear ; this is the systolic pressure. As the pressure is lowered the sounds become louder and louder and then abruptly die away. The point at which the loud sounds begin abruptly to die away is taken as diastolic pressure, for below this point the pressure of the cuff has failed to deform the artery. It is to be noted that while this index is probably a reliable measure of diastolic pressure in most subjects, patients are occasionally encountered in whom the sounds continue to be heard when the pressure

in the cuff is reduced to zero ; in these cases this index of the diastolic pressure is clearly unreliable.

(3) **By the Oscillometer.** In some forms of the instrument the bag is connected to a high-frequency diaphragm type of pressure gauge. The oscillations of the diaphragm record the volume changes of the main artery of the limb transmitted to the air in the cuff. When the pressure in the cuff is slowly reduced, the point at which the oscillations of the manometer first increase is taken as systolic, the point of maximum oscillation as diastolic pressure.

The auscultatory method is that most commonly used in England and the values correspond most closely with direct manometric measurements. The oscillatory method gives values for the systolic and diastolic pressures that are usually 5 to 10 mm. Hg higher than those obtained by the auscultatory method. Systolic pressures obtained by palpation are in experienced hands in fairly close agreement with those obtained by auscultation ; owing, however, to the difficulty of feeling the first weak pulse beat, the values particularly in inexperienced hands are frequently some 5 to 10 mm. lower.

The Capillary Pressure. Accurate estimation of the capillary pressure in man is extremely difficult. One method is to seal on to the skin a small glass chamber open on the side next the skin ; its interior is connected to a manometer and source of air pressure. The pressure of air necessary to cause a capillary loop (observed microscopically) to disappear is taken as the capillary pressure. This method gives variable results and is less reliable than the following direct method. A finger is immobilised in a bed of plasticine and the cuticle is shaved off from the base of the nail. If a drop of glycerine is placed on the skin, the capillary loops can easily be seen with a binocular microscope and surface illumination. By means of a micro-manipulator a fine glass micro-pipette containing physiological saline and sodium citrate solution is introduced into one of the capillaries, and blood allowed to enter its orifice. The pressure at which the blood neither enters nor leaves the pipette but oscillates with each heart beat is the mean capillary pressure. Owing to difficulties of fixation and of observing sufficiently large capillaries, the base of the nail is as yet the only place where the method is practicable in man.

The Venous Pressure. The veins offer little frictional resistance, and the blood flows through them with only a small fall of pressure. The venous pressure at some distance from the heart is thus very close to that in the superior vena cava. The most accurate and direct method of determining the venous pressure in man is to introduce into the median basilic vein at the elbow a wide needle connected with a manometer containing a solution of sodium citrate. The solution is allowed to flow into the vein until the meniscus shows small respiratory oscillations about a fairly constant mean value. The height of this meniscus gives the venous pressure in the vein at the point of measurement. To

obtain a gauge of general venous pressure such values must be corrected for the difference in level between the vein punctured and the heart, since the pressure in the veins, as in all the vessels, is affected by gravitational forces. It is not easy to say precisely at what level the heart is in man, and so the pressures are usually referred to an easily accessible structure bearing a fairly constant relation to the heart, the junction of the manubrium with the body of the sternum (angle of Louis).

In recumbent healthy subjects when the arm lies at or below the level of the heart, the meniscus in the venous manometer comes to rest at the same level as the angle of Louis, or a centimetre or two below it. Relative to the angle of Louis the venous pressure in health is thus 0 to $-$ 2 cm. H_2O.

Now the veins, being wide lax vessels, are distended when the pressure of their contained blood is greater than that of the atmosphere. When the venous pressure is a little below that of the atmosphere the

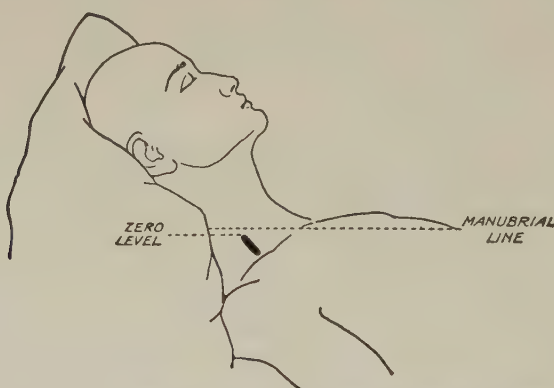


Fig. 2. 3. Measurement of the Venous Pressure in Man.

The subject lies down with his head supported by pillows ; the jugular vein is seen distended with blood up to a point which usually lies just below the level of the notch of the *manubrium sterni*. (Lewis.)

veins collapse. If therefore in the recumbent subject we trace a superficial vein, such as the external jugular, from a point well below the level of the manubrium sterni to a point above it, we see that for the lower part of its course the vein is distended, and then at some higher point it collapses and ceases to be visible (Fig. 2. 3). At the junction of distended and collapsed vein, pulsations synchronous with the heart beat and respiration will be observed. From what has been said it is clear that the junction of distended and collapsed vein should give the level at which the venous pressure is equal to that of the atmosphere, and in fact it is found that when the venous pressure is measured manometrically, the meniscus lies at the same level as that to which the jugular veins are distended. The point of collapse of the jugular veins is therefore used clinically to measure the venous pressure in man. In so doing it is essential to ensure that the point at which the vein ceases to be visible is not simply the point at which it plunges deeply into the

neck ; this may be ascertained by noting that the vein fills to a higher level when it is obstructed below by a finger.

In resting man the venous inflow to the heart is small ; the healthy heart requires but a small distending pressure to expel this inflow and maintains the venous pressure low. The failing heart, which is working at the limit of its capacity, requires a venous pressure that is several centimetres (2 to 10 cm. H_2O) higher to accomplish its task ; and under such circumstances the venous pressure is raised. It may be readily understood, therefore, that in the resting subject a rise of venous pressure above its normal value is the most usual and most important sign of failure of the heart. For this reason measurement of the venous pressure is of unusual importance clinically.

Normal values for the blood pressure in resting man at various parts of the vascular circuit are as follows :—

Axillary artery	systolic	115 mm. Hg
	diastolic	70 ,,
	(pulse pressure	45 ,,)
Capillary of nail fold	arterial limb	32 ,,
	summit of loop	20 ,,
Superior vena cava		0 to — 2 ,,

Of these values the venous pressure is the most constant, rarely varying by more than 2 mm. Hg (3 cm. H_2O) from the mean ; the capillary pressure shows slightly greater variations (5 mm. Hg) and the arterial pressure considerably greater fluctuations. Thus in an individual examined at rest on several occasions, the systolic pressure may vary from 110 to 130 mm. Hg and the diastolic pressure from 60 to 80 mm. Hg ; psychical factors are amongst the more important causes of these variations. Average values for the systolic blood pressure rise from 80 to 100 mm. at the age of five, to 115 mm. at the age of puberty and to 140 mm. Hg at the age of sixty.

Fig. 2. 4 shows the blood pressure determined directly in various parts of the vascular circuit of the guinea pig by the introduction of a micropipette into the appropriate vessel. It will be seen that there is no appreciable fall of pressure until the smaller arteries are approached, then the pressure falls rapidly until the capillaries are reached, when the fall becomes more gradual. The fall of pressure as the blood traverses the veins is very small. The figure also shows that the pressure in the arterial limb of the capillary is higher, and in the venous limb lower, than the colloid osmotic pressure of the blood ; fluid thus tends to pass out of the blood at one end of the capillary and to be absorbed at the other (see Chapter 8). Owing to the large and variable resistance offered by the arterioles, the capillary pressure is largely independent of the general arterial pressure, and follows more closely the venous pressure, since there is little resistance to the flow from capillaries to veins. In an active organ, as we shall see, there is considerable arterial dilatation, and increase in rate of blood-flow. The capillaries also dilate, and those previously closed, open up.

The Effect of Gravity. The above values of the blood pressure are all given for vessels lying at the same level as the heart. When the

vessels lie below this level, then to the pressure which is imparted to the blood by the heart must be added the pressure due to gravity, that is the pressure exerted by a column of blood equal in height to the vertical distance of the vessels examined from the heart. This relationship holds good for all the vessels of the perfectly flaccid limb; but in dependent limbs the venous pressure is reduced by the repeated movements which are usual during active life. Even the smallest movements empty the veins which, owing to the action of the valves, can only fill up from below. Thus in health the venous pressure in dependent limbs rarely rises very much above that of the atmosphere. This is important

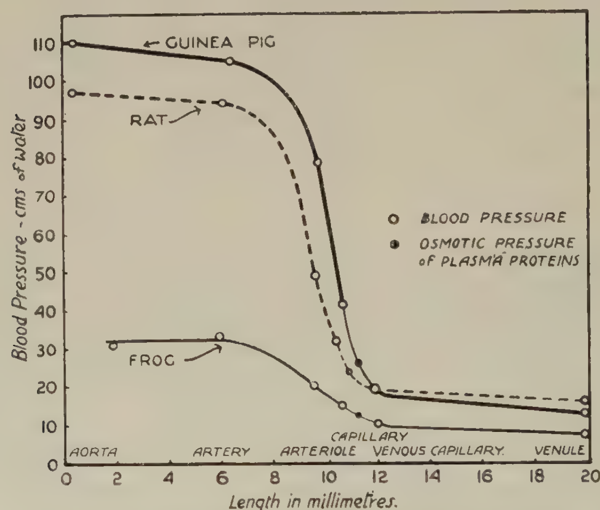


FIG. 2.4. Shows the **mean blood pressure** (circles) determined by inserting a micropipette into different parts of the mesenteric circulation of the guinea-pig, rat and frog. In each curve the thin line represents the fall of pressure along the capillaries. The dots indicate determinations of the osmotic pressure of the plasma proteins in each species. (After Landis.)

because if the venous pressure becomes greatly raised, then the capillary pressure is correspondingly raised and fluid tends to pass out of the capillaries and waterlog the tissue spaces. If the leg of a normal person is allowed to hang down for several hours without movement, it thus becomes œdematous. When the veins of the leg are distended by disease (varicose veins), the valves become incompetent, and after a day's work the feet become swollen because of the raised capillary pressure.

Factors Determining the Arterial Blood Pressure

As we shall see later, the maintenance of an adequate supply of blood to the brain and to the heart is intimately dependent on the level of the general arterial pressure. This depends on the cardiac output and on the peripheral resistance; the resistance offered by the vessels

being determined by their diameter and by the viscosity of the blood. In health the cardiac output is mainly determined by the venous return to the heart, and this largely depends on the relationship of the total blood volume to the capacity of the circulation.

The Arterial Blood Pressure in Man. The mean pressure is generally taken as the diastolic pressure plus one-third of the pulse pressure. The relation between the cardiac output (CO), the arterial pressure (BP), and the total peripheral resistance (TPR) is defined by the relation : $CO = BP/TPR$. The fluctuation of the arterial pressure above and below the mean, that is to say the pulse pressure, depends chiefly upon how much blood is thrust into, and escapes from, the arterial tree between successive heart beats, and this depends chiefly upon the stroke volume. Increase in cardiac output by itself increases both mean pressure and stroke volume, and therefore raises systolic, diastolic and pulse pressures ; increase in peripheral resistance by itself increases the mean pressure without altering the stroke volume and therefore increases the systolic and diastolic pressures without much altering the pulse pressure ; increase in the pulse rate by itself does not affect the mean pressure, but reduces the stroke volume ; the pulse pressure, therefore, diminishes and the systolic and diastolic pressures approach the mean pressure. These relations, though useful as will be seen later, are only approximate ; and attempts to calculate the cardiac output accurately from formulæ embodying the systolic and diastolic pressures and the heart rate have not been very successful.

Factors Determining the Venous Return to the Heart. At the pressures normally ruling in the vascular system, the veins are the most distensible vessels, and it is to be anticipated that when blood is added to the circulation most of it will be accommodated in the veins and thus increase the venous return to the heart. The same result should occur if the vessels in a given territory constrict so that blood drains out of them into the great veins. The venous return to the heart should be increased, therefore, either if the total blood volume increases, or if the capacity of the circulation is decreased. This can be shown very simply in a preparation in which the heart propels blood through a closed system analogous to that formed by the intact vessels ; injection of blood into this system greatly increases the output of the heart ; withdrawal of blood diminishes the output. The capacity of the circulation is determined by the general tonus (state of contraction) of the arterioles, capillaries and veins, but there are a number of situations in which blood is particularly liable to accumulate and which merit the term **blood reservoirs.** Of these the spleen is peculiar in that the blood is accommodated and concentrated by loss of plasma, in the spleen pulp, a bypass of the circulation ; in the dog the spleen can accommodate one-fifth of the total volume of blood. The other reservoirs, the liver and portal system, the sub-papillary venous plexus of the skin, and the great veins are, unlike the spleen, part of the general circulation, but are, like the spleen, capable of considerable variations in capacity ; the liver and portal system (including the spleen) at rest

contain about one-quarter to one-third of the total volume of blood in the cat and dog, the skin very much less. The capacities of these structures in man are not accurately known, but it is probable that the skin is relatively more and the spleen relatively less capacious than in the cat and dog.

The Control of the Peripheral Vessels

From the data already given for the blood pressure in different parts of the vascular circuit (Fig. 2. 4, p. 40) it will be seen that the main fall of pressure occurs in the small arteries and arterioles. The peripheral resistance is thus chiefly constituted by these vessels and its magnitude is dependent upon the strength of contraction of their smooth muscle coats. The state of contraction, or " tone " of the arterioles, may be modified for one of two purposes : to fulfil local metabolic requirements or to safeguard the circulation to the brain.

Local Control. The arterioles are relaxed, or dilated, locally in any organ or tissue when its activity increases and there is a greater demand for oxygen. The most important of these organs are the heart, the skeletal muscles and the digestive glands. Such an adjustment of the circulation is brought about chiefly by the production in the active tissues of vaso-dilator substances which act directly on the arterioles ; the blood flow is increased and a greater supply of oxygen is made available. This will be discussed again later.

Reflex Control. The arterioles in many parts of the body, on the other hand, may become constricted, or allowed to dilate, so adjusting the total peripheral resistance as to maintain the arterial pressure and thus the blood flow to the brain. A simple vaso-dilatation in any organ will lower the peripheral resistance and hence the arterial pressure. It is not surprising, therefore, that mechanisms exist which ensure that when vaso-dilatation occurs in one part of the body, this is compensated by vaso-constriction elsewhere. This control is initiated by receptors in the walls of certain arteries which are sensitive to the level of the arterial pressure and are thus called *baroreceptors*. The most important of these are situated : (*a*) in the arch of the aorta, giving rise to the " depressor reflex "; and (*b*) in the carotid sinus, the name given to the expansion at the origin of the internal carotid artery. The baroreceptors send nerve impulses to a group of nerve cells known as the " vaso-motor centre " lying in the floor of the fourth ventricle close to the vagus nucleus. These in turn send impulses through the autonomic nerves to the blood vessels.

The Vaso-motor Centre. The maintenance of the general vaso-motor tone is intimately dependent on the integrity of this centre. Section of the hind-brain below the level of these cells leads to generalised vaso-dilatation, and the blood pressure falls from, say, 120 to 80 mm. Hg. After several days, if the animal survives, the blood pressure may rise again almost to its previous level ; destruction of the spinal cord reduces the blood pressure almost to zero. It thus appears that there are also vaso-motor centres in the spinal cord, but in ordinary circumstances

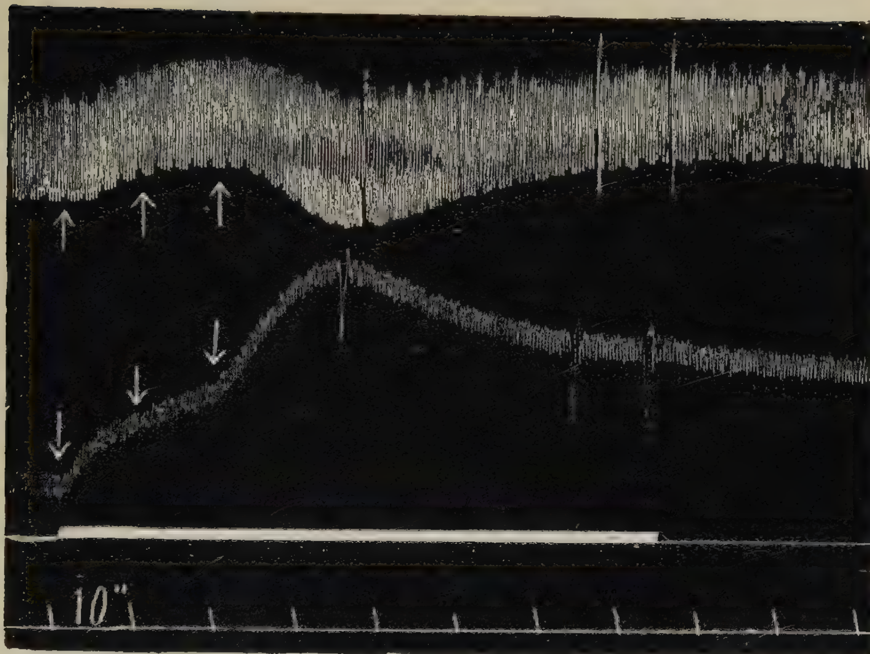


FIG. 2.5. The Effect of **Stimulating the Splanchnic Nerve** on the Arterial Blood Pressure, and on the Output and Volume of the Ventricles.

H., heart volume (a rise in the curve indicates an increase in volume). B.P., arterial blood pressure. S, signal showing duration of stimulation of the splanchnic nerve. T., time marker, showing 10 sec. intervals. Note that the first rise in arterial pressure is associated with an increase in the volume of the heart, owing to the greater power needed in order to expel the blood against the raised pressure, but that the secondary rise, due to the **secretion of adrenaline**, is associated with a decrease in the volume, showing that the heart is beating more forcibly. Note also that the output is increased by the presence of adrenaline, as shown by the increased excursion of the cardiometer record (stroke volume). (From Starling's " Principles of Human Physiology.")

these are mainly controlled by that of the hind-brain (medulla) ; when this is destroyed, then the spinal centres gradually take over control. Electric stimulation of the vaso-motor centre, on the other hand, leads to generalised vaso-constriction and a rise of blood pressure. The normal activity of the centre in maintaining general vaso-motor tone is, as we shall see, profoundly modified by the influence of the cerebral hemispheres, as well as by afferent impulses from the baroreceptors, and by chemical stimuli, the centre being stimulated by carbon dioxide and by inadequate oxygen supply.

The role of carbon dioxide in determining the activity of the cells of the vaso-motor centres is well shown when the gas is excessively removed from the blood by over-ventilating the lungs. If the lungs of an anæsthetised cat are artificially over-ventilated with air, the blood pressure may fall in two minutes from 140 to 40 mm. Hg. If air containing 5 per cent. carbon dioxide is substituted for the ordinary air previously used, the rate of ventilation remaining the same, the blood pressure returns to its original level.

After a cerebral hæmorrhage the blood pressure may rise to 150 mm. Hg

and the pulse is slowed. The rise of blood pressure is usually attributed to anæmia of the vaso-motor centre, arising from compression of the brain by the hæmorrhage ; in this way the cells of the centre would be stimulated by oxygen lack and the accumulation of metabolites.

The Vaso-motor Nerves. The nervous control of the blood vessels is chiefly effected through the sympathetic branches of the autonomic nervous system (Chapter 15). The action of the vaso-motor nerves was discovered by Claude Bernard, who found in the rabbit that when the cervical sympathetic chain was divided on one side, the ear on that side became flushed and warm, remaining so for a considerable time. Conversely, stimulation of the cervical sympathetic produces pallor

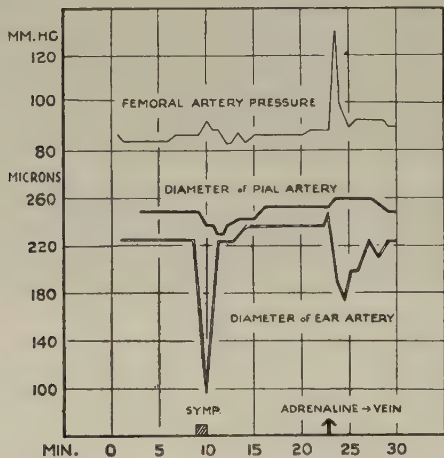


FIG. 2. 6. The lower two curves record in microns the diameters of an artery of the pia mater of the brain, and of an ear artery, observed microscopically in a cat. The pial artery was seen through a glass window screwed into the skull. At the signal " symp." the cervical sympathetic trunk was stimulated and produced a pronounced constriction of the ear artery but only a slight narrowing of the pial artery. At the next signal 0·01 mg. adrenaline injected into a vein produced a marked constriction of the ear artery and a small dilatation of the pial artery which is to be attributed to a passive effect of the coincident rise of blood pressure. (Forbes, Firnley and Mason.)

and coldness of the corresponding ear. We now know that if the appropriate branches of the sympathetic nerves are stimulated under suitable conditions, then the blood vessels in all parts of the body (except perhaps the heart) constrict. The degree of constriction, however, varies in different organs. Thus stimulation of the appropriate sympathetic fibres produces intense narrowing of the vessels (vaso-constriction) of the skin and alimentary canal (Fig. 2. 5), but only slight narrowing in those of the brain and lungs (Fig. 2. 6). The vessels constricted by stimulation of the sympathetic are particularly the arteries and arterioles, to a less extent the capillaries and the veins. Section of the sympathetic nerves supplying an organ leads to an increase in the blood-flow through it, and it thus appears that normally there is a steady stream of constrictor

impulses passing along these nerves to the blood vessels. Further, since in the majority of organs the vessels cease to participate in the vaso-motor reflexes, shortly to be described, after their sympathetic fibres have been cut, it seems that changes in vascular calibre of vaso-motor origin are chiefly determined by an increase or decrease (inhibition) of sympathetic vaso-constrictor impulses.

Sympathectomy. Division of the sympathetic fibres supplying a limb, either by removal of the appropriate sympathetic ganglia or section of the preganglionic sympathetic fibres, is followed by vasodilatation, by loss of vascular responses to change of body temperature, and loss of reflex sweating and pilomotor response. The vasodilatation chiefly affects the skin, muscle blood-flow being little altered. With the progress of time, the vasodilatation subsides and the vessels are found to be abnormally sensitive to adrenaline, histamine and other vaso-active substances. Complete sympathectomy, by removal in separate stages of the chain of sympathetic ganglia on both sides, has been performed in the cat and dog. The completely sympathectomised cat is sluggish, and very susceptible to exposure to cold, to oxygen lack, and to hæmorrhage. The sympathectomised dog is normally active and its arterial pressure at rest is little below normal. It is evident that while in the normal animal the vasomotor nerves are extremely important in regulating the circulation, yet other and probably chemical mechanisms can in the dog take over much of this function (see Chapter 15).

Reciprocal Innervation and Vasodilator Nerves. It was thought at one time that the blood vessels generally were supplied by both sets of autonomic nerves—sympathetic (constrictor) and parasympathetic (dilator)—the two sets acting reciprocally. Thus an increase in flow would be brought about by a decrease in impulse frequency in sympathetic fibres and by a concomitant increase in impulse frequency in parasympathetic fibres. Later studies showed that direct parasympathetic vaso-dilator fibres probably occur only in the nervi erigentes supplying the erectile tissue of the genital organs (Chapter 10). The vaso-dilatation that occurs in the salivary glands during parasympathetic nerve stimulation is not brought about by true parasympathetic vaso-dilator fibres, but by the action of vaso-dilator substances formed during stimulation of the secretory cells. Skeletal muscle vessels in the cat are supplied by sympathetic vaso-dilator fibres ; but these do not act reciprocally with the vaso-constrictor fibres, but quite independently.

The Depressor Reflex. The baroreceptors in the arch of the aorta send nerve fibres to the hind-brain in the trunk of the vagus (as in man) or as a separate " depressor " nerve (rabbit). Stimulation of these fibres produces a slowing of the heart and fall of blood pressure. The slowing of the heart is largely but not entirely abolished by previous section of the vagi ; it is thus due mainly to a reflex augmentation of vagal tone, and partly to a reflex inhibition of sympathetic tone. The fall of blood pressure is independent of slowing of the heart and is due to a vaso-dilatation that affects all the organs of the body except perhaps the brain (Fig. 2. 7). Thus stimulation of the depressor nerve produces an increase in the volume of a limb or of a loop of intestine, and an increase in blood-flow from the submaxillary gland. The physiological stimulus exciting the depressor reflex is a rise of pressure in the arch of the aorta.

The Carotid Sinus Reflex. It has long been known that in man pressure over the bifurcation of the common carotid artery produces a

sensation of faintness accompanied by slowing of the pulse and fall of blood pressure. This effect has been shown to be due to stimulation of baroreceptors lying under the adventitia of the carotid sinus. If the sinus is compressed, or if the pressure of the blood inside is raised, or if the sensory nerve to it (a branch of the glosso-pharyngeal) is stimulated, the blood pressure falls and the heart slows, changes produced reflexly in a manner similar to those of the depressor reflex. Conversely, if the common carotid artery is compressed so as to produce a fall of pressure within the sinus, the heart accelerates and the blood pressure rises. The paths followed by both this and the depressor reflex are very similar. The afferent impulses entering the hind-brain through the glosso-pharyngeal and vagus nerves reach the cardio-inhibitory,

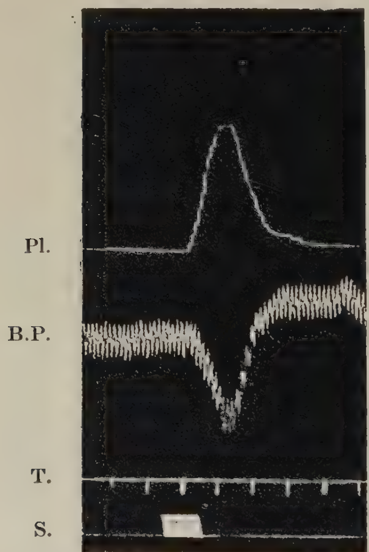


Fig. 2. 7. **The Depressor Reflex, producing Vaso-dilatation and a Fall in Arterial Pressure.**

From above downwards: Volume (plethysmograph) of a loop of intestine, arterial blood pressure, time in 10 sec. intervals, signal showing period of stimulation of the central end of the vagus (containing the depressor fibres). (W. M. Bayliss.)

cardio-accelerator and vaso-motor centres lying close by in the medulla. The impulses sent out by these centres through the vagus and sympathetic nerves are thus modified in the way described.

If the carotid sinus and depressor nerves are cut, the blood pressure and pulse rate rise permanently. Blood pressure and pulse rate in man rise if the carotid sinus nerves are blocked by local anæsthesia (Fig. 2. 8). In normal life, therefore, the constancy of the blood pressure and pulse rate is due largely to impulses ascending these nerves ; any variation in blood pressure produces an inhibition or augmentation of these impulses, and so reflexly initiates changes which restore the blood pressure to its normal level.

When the arterial blood pressure falls, these reflexes produce a general increase in vaso-motor tone. The vessels that actually constrict vary according to circumstances, and appear to be those in which it is unimportant to maintain a large flow. Thus the vessels of the brain are

little constricted and those of the heart are dilated by sympathetic stimulation, and do not appear to participate in the general vasoconstriction evoked through the carotid sinus and depressor reflexes ; in the warm animal the vessels of the skin are dilated and constrict little in the reflex ; in the cool animal, the narrow skin vessels constrict very markedly. These pressor effects are reinforced by the simultaneous release of adrenaline and noradrenaline. The two reflexes are thus of extreme importance in maintaining the distribution of blood to the tissues according to their needs ; at the same time they prevent undue strain on the heart by keeping the blood pressure within convenient limits.

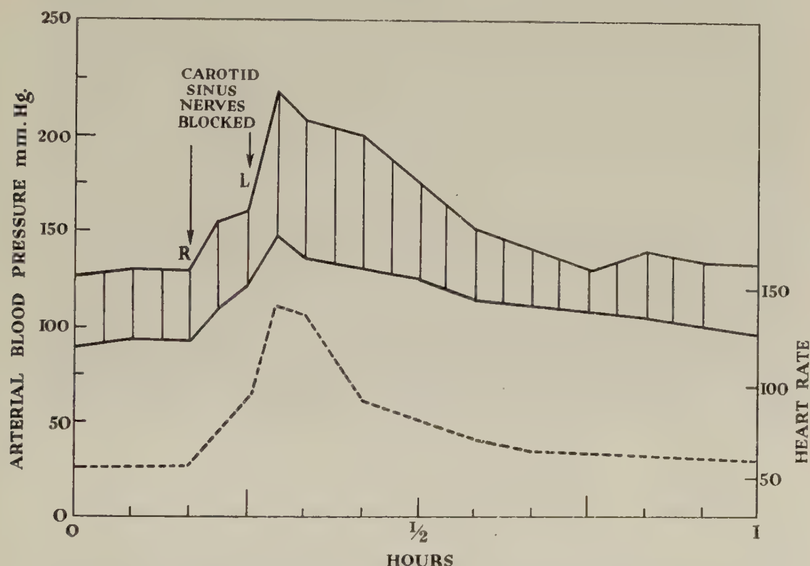


FIG. 2. 8. **Arterial Blood Pressure and Heart Rate** in the Human Subject before, during, and after **blocking the carotid sinus nerves** with local anæsthetic. After release of the vasomotor centre from the inhibitory impulses of these baroreceptors, the arterial pressure rose to 224/148 mm. Hg (systolic/diastolic), and the heart rate increased to 140 beats per minute. (After Lampen, Kedzi and Kaufmann.)

The methods used by Heymans and his co-workers in investigating the functions of the carotid sinus are interesting as an example of physiological technique (Fig. 2. 9). One or both carotid arteries of one dog B are perfused with blood from a second dog A (cross-circulation). The head of dog B is completely severed from its trunk, except for the spinal cord and vagus nerves. On raising the arterial pressure in dog A and thus in the carotid sinuses of dog B, the blood pressure in the trunk of dog B falls and the heart rate diminishes ; the opposite changes occur when the blood pressure of dog A is lowered (Fig. 2. 10). If now the suprarenal vein of dog B is anastomosed with the internal jugular vein of a third dog C (Fig. 2. 9), then a fall of blood pressure in dog A produces, in addition to the effects mentioned, a contraction of the spleen of dog C (Fig. 2. 10). Thus a fall of pressure in the carotid sinus of dog B leads in this dog to an increased secretion of adrenaline, as shown

by the effect of blood from its suprarenal vein on the spleen of C. The effects are abolished by denervating the carotid sinuses of dog B.

Responses of the General Circulation

In the course of life, a man is likely to be subjected to various conditions, resulting from activity or accident, which will affect the circulatory system as a whole. The circulation will respond to these conditions in various appropriate ways, which we will now describe.

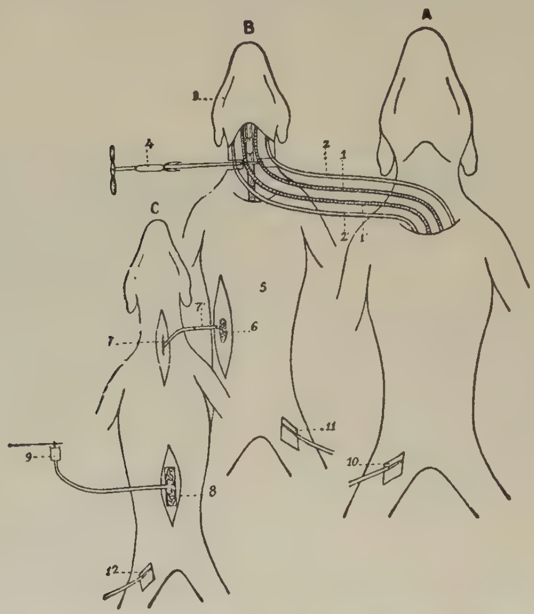


FIG. 2. 9. Arrangement of Animals in a **Cross-Circulation Experiment.**

The head of dog B (3) is perfused from dog A by anastomosis of the carotid arteries and also of the jugular veins (1, 1′, 2, 2′), the vertebral arteries and veins being tied, the muscles of the neck divided between ligatures, and the rest of the tissues compressed round the vertebral column by the *écraseur* of Chassaignac (4). Blood from the suprarenal of B (6) is led into one of the jugular veins (7) of C, which has been adrenalectomised ; the spleen of C (8) is enclosed in a plethysmograph, and contracts whenever adrenaline is secreted by the suprarenals of B. The arterial blood pressures of the three dogs are recorded from the femoral arteries 10, 11, 12. (From C. Heymans.)

Posture. When a recumbent subject stands up, blood pools in the veins of the lower part of the body, there is less in the great veins to fill the heart and cardiac output diminishes. Arterial blood pressure tends to fall, but the depressor and carotid sinus reflexes respond at once and prevent any significant alteration.

Exercise. When exercise is imminent, the higher centres of the brain prepare the circulation to some extent. Vagal inhibition of the heart subsides, the heart beats more rapidly and the force of its contractions

is increased by sympathetic stimulation and by the action of adrenaline. At the beginning of exercise, heart rate and cardiac output rise rapidly and in about three minutes will have reached levels related to the severity of the exercise. As is shown in Table 2. 2 the increase in output is due partly to increase in the frequency of the heart and partly to increase in the amount expelled by each beat (stroke volume). X-ray pictures of the human heart show no change in the diastolic volume,

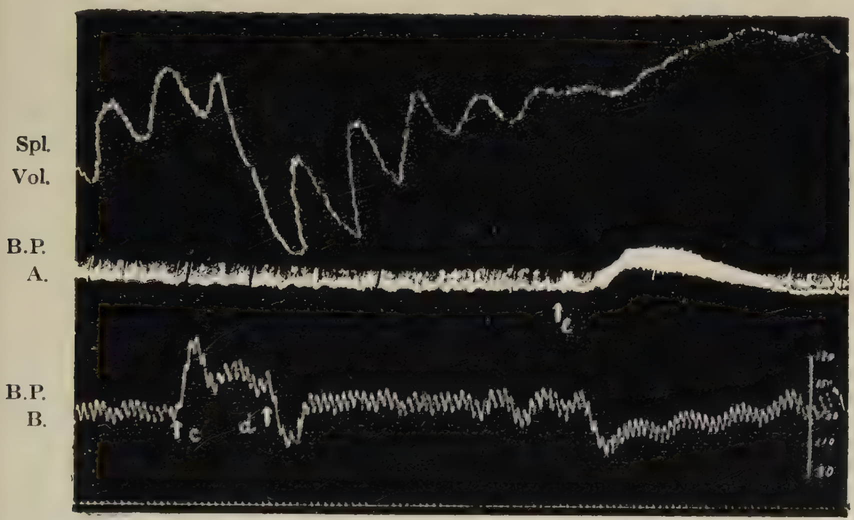


FIG. 2. 10. **The Regulation of the Arterial Blood Pressure.** (Cross-circulation experiment, see Fig. 2. 9.)

From above downwards : Volume of the spleen of dog C ; arterial pressure of dog A (perfusing the carotid sinus of dog B), arterial pressure of dog B.

At *c* the pressure in the carotid sinus of dog B was reduced by partially clamping the inter-connecting artery, and at *d* the pressure was returned to the initial value. A **fall in pressure in the carotid sinus** leads to a reflex rise in pressure in the rest of the body, and a reflex secretion of adrenaline, as shown by the contraction of the spleen of dog C ; a rise in pressure in the carotid sinus has the reverse effect. At *e* 0·1 mg. of adrenaline was injected into the circulation of the perfusing dog, A, **raising** its arterial pressure, and hence, also, **the pressure in the carotid sinus** of dog B. This resulted in a reflex fall in the arterial pressure of B, and an inhibition of the secretion of adrenaline, as shown by the dilatation of the spleen of dog C. (From C. Heymans.)

and the increase in stroke volume results from the ventricles emptying more completely in systole. The diastolic volume of the ventricle of the resting untrained subject in the upright position is about 90 ml. and the stroke volume 45 ml. leaving 45 ml., behind—the residual volume. In very severe exercise, the ventricle empties almost completely and the stroke volume is nearly 90 ml. In trained subjects the heart hypertrophies and enlarges and in severe exercise the ventricle can discharge up to 200 ml. per beat (Table 2. 2).

The venous return, which in the steady state of course must equal

TABLE 2. 2

The Effect of Exercise on the Output of the Heart (after Christiansen)

Subject.	Work performed. Kg-metres per min.	Oxygen consumption. Litres per min.	Pulse rate per min.	Cardiac output. Litres per min.	Output per beat. c.c.
Untrained female	0	0·24	·77	4·6	60
	600	1·57	131	14·5	111
	720	1·79	145	17·4	120
	840	2·05	159	19·0	120
	960	2·45	168	23·8	142
Trained male . . .	0	0·25	70	4·2	60
	720	1·93	118	16·5	140
	960	2·22	140	20·6	147
	1,200	2·83	174	23·0	132
	1,440	3·26	180	26·9	149
	1,680	3·94	179	37·3	208

the cardiac output, is increased proportionally to the work done. Venous pressure, measured through a catheter in the great veins near the heart or in the right auricle, remains almost constant. From the experiments on the heart-lung preparation, it will be recalled that increase in the rate and force of contraction of the heart increases the output at a given value of the venous pressure (Fig. 1. 12, p. 22). In a given time, a much greater volume of blood can enter the heart because the duration of the diastolic pause is increased at the expense of the systolic pause and because the ventricle relaxes more rapidly and offers less resistance to the entering blood. Owing to the increase in the force of contraction, also, the ventricle empties more rapidly and completely from a given diastolic volume.

Although in exercise the venous pressure remains constant, there is probably an increase in the " effective filling pressure ." This is the difference between the pressures outside and inside the heart—that is between the pressure outside the chest, and the intra-pleural (intra-thoracic) pressure. Since the intra-pleural pressure is decreased owing to the increased respiratory movements, there is probably a corresponding increase in the effective filling pressure (see below, p. 99).

We must now consider the cause of the increase in venous return. In the resting dog, stimulation of the heart causes a fall in venous pressure and only a small increase in output of the heart. It is therefore clear that in exercise the cause of the increase in venous return is to be sought in the peripheral vascular system. If cardiac output increases six-fold, the peripheral resistance must decrease to about one-sixth of its resting value if the mean arterial pressure is to rise only slightly. The decrease in resistance is mainly in the vessels of the skeletal muscles,

and to a lesser extent in those of the skin. If cardiac output increases from 5 to 30 litres per minute, then the flow through the muscle vessels will increase from 1 to 22 or 23 litres per minute, and that through the skin vessels will increase from a fraction of a litre per minute to 2 or 3 litres per minute. The increase in flow through the skin helps to prevent the body temperature from rising (Chapter 20). In spite of this enormous decrease in the resistance of the vessels in the muscles and skin of the limbs, there is little or no increase in the limb volume. This is due to the action of the " muscle pump " (Fig. 2. 11). When we are at rest, a good deal of blood is contained in the veins of the limbs. In exercise, muscular contraction squeezes these veins and propels the blood towards the heart, the venous valves preventing reflux from the heart. Thus the increase in the amount of blood in the dilated arterioles is more than redressed by the decrease in the amount of blood

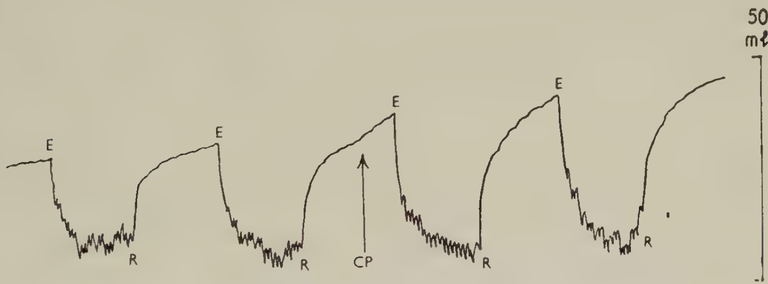


Fig. 2. 11. Shrinkage of the Calf of the Leg due to the action of the " Muscle Pump."

Calf volume changes recorded with plethysmograph: downward movement of the record denotes shrinkage. E, rhythmic exercise of calf muscle for 10 sec.; R, rest for 10 sec.; CP, pneumatic cuff applied just above the knee and inflated to 90 mm. Hg; this was maintained to the end of recording. (Barcroft and Dornhorst.)

in the limb veins. Vaso-constriction occurs in the viscera, and the volume of blood within their vessels decreases, with a corresponding increase in the amount of blood in the active circulation, for which accommodation must be found. Owing to the " muscle pump " and to this redistribution of the blood, the central parts of the venous system contain more blood than they do in the resting subject. This tends to raise the venous pressure, and thus to increase venous return to the heart. But the stimulated heart offers less resistance to the entering stream of blood, and this will tend to lower the venous pressure (as indicated in Fig. 1. 12). As a result, there is a large increase in the venous return without any substantial rise in pressure in the neighbourhood of the right auricle.

The following imaginary experiment may help the reader to visualise the part played by the peripheral vascular system in increasing the venous return during exercise. Suppose that the peripheral vascular system of a resting animal is being perfused by a pump. Blood is

pumped from the venous side through an oxygenator into the aorta. The output of the pump is regulated by the experimenter in such a way as to keep constant the pressure in the great veins (venous pressure) at the normal resting level. Now suppose that the animal begins severe exercise. Vaso-dilatation begins in the muscles and the arterial pressure falls. The " muscle pump " and the redistribution of blood from the viscera fill up the great veins and the venous pressure rises. The operator, watching the venous pressure, adjusts the pump so as to increase its output and bring back the venous pressure to the normal resting level. Several such adjustments will be necessary, increasing the output and work of the heart, until the redistribution of blood from the viscera is complete and steady conditions are established. Finally, the venous pressure will be constant at the normal resting level, the output and work done by the pump will be increased about six-fold, the mean arterial pressure will be the same as it was initially, or slightly raised, and the peripheral resistance will have fallen to one-sixth of its initial value. In exercise, the heart, quickened and strengthened, automatically adjusts its output so as to maintain the venous pressure constant.

The effect of redistribution of the blood on cardiac output may be illustrated by the following experiment. When in an anæsthetised dog with the central nervous system destroyed, the descending thoracic aorta is clamped, the arterial pressure rises and, surprisingly enough, the cardiac output *increases.* Although the arterial supply to two-thirds of the systemic circulation is cut off, and in spite of the great increase in the peripheral resistance, the rate of circulation round the remaining one-third, comprising the vessels of the head and upper part of the body, actually increases more than three-fold, so that it comes to exceed the original total cardiac output. The explanation is as follows. After the descending aorta is clamped, the distended vascular bed in the lower part of the body gradually collapses, discharging about 50 ml. of blood. Accommodation must be found for this 50 ml. of extra blood in the active circulation, and most of it accumulates in the veins of the head and neck. Distension of these veins raises the pressure in the superior vena cava so much that venous return and inflow into the heart from the upper part of the body finally exceed the original venous return and inflow from the whole body.

Oxygen Lack. The effects of oxygen lack (anoxia, or better **hypoxia,** since oxygen is not completely absent) have been studied extensively owing to their importance in flying. When the oxygen pressure in the inspired air is reduced to one-half, the heart beats much faster (tachycardia, p. 27 above) and the cardiac output and systolic pressure are increased. Reduction of the oxygen pressure to one-third of its normal value causes loss of consciousness ; in most subjects, the " non-fainters," the circulatory changes just described are accentuated ; in the others, the " fainters," vaso-vagal fainting occurs (see below).

Hæmorrhage. If blood is lost, and the volume in circulation is reduced, it is to be expected that the veins would be less well filled and that the output of the heart and the arterial pressure would fall. This does occur, particularly if the hæmorrhage is severe, but the changes are smaller than might be expected owing to the existence of compen-

sating mechanisms. Compensation is absent after section of the spinal cord in the neck and is thus effected through the central nervous system. If, in an anæsthetised dog, the blood flow through the gut, the limbs, the liver and the spleen are simultaneously measured, it is found that when, say, one-tenth of the blood volume is removed, the flow through the gut and the limbs is reduced, showing that the vessels constrict in these areas. The arterial pressure does not fall, but is maintained by an increase in peripheral resistance brought about by a reduced activity of the depressor and carotid sinus reflexes. On the other hand, the outflows in the veins from the liver and spleen are found to be temporarily increased and exceed the inflows through the arteries : the blood content of these organs is largely expelled into the great veins, partly compensates for that removed, and lessens the fall in cardiac output. In the rabbit and the cat another compensatory mechanism quickly comes into action : fluid is absorbed from the interstitial spaces through the capillary walls, diluting the blood and partially restoring its volume. In the dog and man dilution of the blood is slower, and in man is not complete until 24 to 72 hours after removing 1 litre of blood. But if the quantity of blood lost is sufficient to reduce the cardiac output to 30 to 50 per cent. below normal, the arterial pressure falls. After profuse hæmorrhage in man, the fall may be profound enough to produce loss of consciousness through anæmia of the brain ; there may be complete recovery, nevertheless.

When the circulation in the anæsthetised dog is so disturbed by bleeding that the mean arterial pressure has fallen to 30 or 40 mm. Hg and remained so for six hours, a return of all the blood removed does not lead to a lasting recovery of the circulation (irreversible shock). If, after severe hæmorrhage, the irreversible state is to be prevented, prompt measures must be taken to restore the depleted blood volume. This can be most effectively done by transfusing fresh or stored blood from a healthy donor of the same or a compatible blood group (see Chapter 21). In an emergency, stored plasma or serum may also be used, but in any case the amount transfused must be adequate to restore arterial pressure to normal. Saline is useless for the purpose for, having no colloid osmotic pressure, it quickly passes out into the tissue spaces. It is obviously desirable, also, to know the volume of blood in circulation. Methods for measuring this are described in Chapter 8 (p. 226). In most animals, the blood normally makes up from 5 to 8 per cent. of the body weight, there being about 3 to 5 litres in an average man, and 500 to 1,500 ml. in an average dog of 6 to 20 kg. weight.

Peripheral Circulatory Failure and Shock. A condition resembling that seen after frank hæmorrhage in which the blood pressure falls, and in which the reduction in cardiac output is to be ascribed not to cardiac weakness but to changes in the vessels or circulating blood volume, is described as peripheral circulatory failure or, more loosely, shock. The following are some of the more important examples.

(1) The **vaso-vagal** or **fainting attack** is characterised by low blood pressure, slow pulse, pale cold sweating skin, and sometimes loss of consciousness. The slowing of the heart is effected through the vagus nerves and is

abolished by atropine. The fall of blood pressure is due to vaso-dilatation in voluntary muscle effected through the nerves ; the blood flow through the skin decreases ; the cardiac output may fall. The intense pallor of the skin persists after the arterial pressure has returned to normal, and may be due to the copious secretion of the posterior pituitary which appears to occur during a fainting attack.

(2) **Burns.** In severe burns there is a rapid and profuse loss of plasma into the burned and adjacent tissues. The blood volume falls and the hæmoglobin content of the blood rises, and in severe cases the circulation may fail. Failure is prevented by transfusion of adequate amounts of plasma.

(3) **Wound Shock.** Circulatory failure frequently occurs after extensive wounds, and sometimes without clear evidence of severe blood loss, in which case it has been attributed to vasodilatation from release of a histamine-like substance. Experience in the war of 1939–45 confirmed that in the war of 1914–18, in showing conclusively that wound shock is not a single entity ; but the most important cause of peripheral circulatory failure after wounds is undoubtedly loss of blood, either externally, or into the tissues of the body. In fact the enormous saving of life after wounding in the second war has been due to the provision of proper supplies of blood for early and adequate transfusion (amounts up to 7 litres have been given before and during operation) and to the use of antibacterial agents.

Chemical Regulation of the Circulation

Adrenaline and Noradrenaline. These two substances are produced by activity of the sympathetic nervous system, are secreted by the adrenal gland (Chapter 11), and are released at sympathetic nerve endings in blood vessels and the heart (Chapter 15). In the resting human subject, minute amounts of both substances have been detected in the blood, and may be responsible to a small extent for maintaining vasoconstrictor " tone." During excitement and exercise, activity of the sympathetic nervous system increases and greater amounts of the sympathomimetic amines circulate in the blood. Their effects on the human circulation may now be briefly described.

When adrenaline is infused intravenously in man at about the rate corresponding to maximum secretion of the substance, the subject goes pale owing to constriction of the cutaneous vessels, and he soon becomes aware of his heart beats (palpitation). Fig. 2. 12 shows the response of the general circulation. The heart often beats a little faster owing to stimulation of the pacemaker. There is a considerable rise in the systolic pressure, the mean pressure changes little, and there is a slight fall in the diastolic pressure. Cardiac output is increased (see also Fig. 2. 5). Since the output increases relatively much more than the mean blood pressure, it follows that the peripheral resistance must decrease (see p. 41 above). That is, in physiological doses in man, adrenaline causes an overall peripheral vaso-dilatation. This is because it dilates the splanchnic, skeletal muscular, and coronary vessels more than it constricts those of the skin, kidneys and other organs.

The action of noradrenaline is rather different. The subject pales, but feels no palpitation. Both systolic and diastolic pressures are raised, but, since the cardiac output is decreased, noradrenaline must constrict the peripheral vessels strongly. It is interesting to note that the heart

usually beats more slowly (bradycardia); and this is due to the large rise in arterial pressure and strong stimulation of the baroceptors in the aortic arch and carotid sinuses. Reflex vagal inhibition swamps the rather weak direct excitatory effect of noradrenaline on the pace-maker.

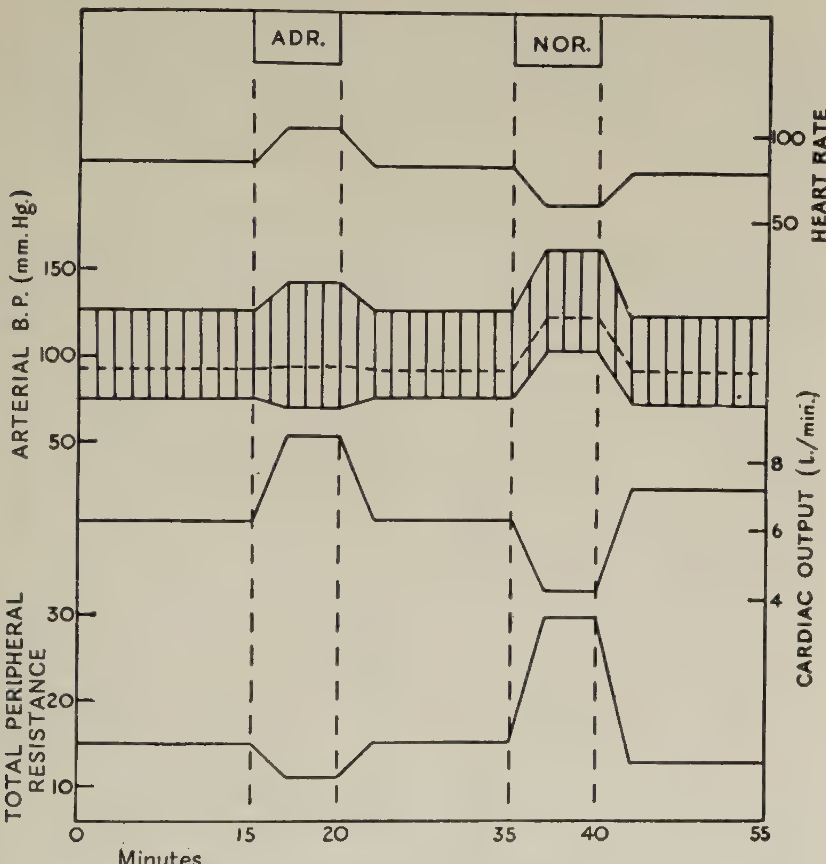


FIG. 2. 12. Diagrammatic representation of the **Effects of Intravenous Infusions of Adrenaline and Noradrenaline** on the heart rate, arterial blood pressure, cardiac output and total peripheral resistance, in man. The infusions were at the rate of 10 micrograms per minute. (*From* Barcroft and Swan, "Sympathetic Control of Human Blood Vessels." Edward Arnold and Co.)

In some medical and surgical emergencies, vaso-constrictor tone is markedly reduced and the arterial pressure may fall dangerously. Normal tone may be temporarily, and sometimes permanently, restored by intravenous infusion of noradrenaline. If necessary, infusion can be continued for several hours. When applied in large doses locally, adrenaline causes strong vaso-constriction. For this reason it is often added to local anæsthetic solutions to localise them near the point of injection and so prolong their effect.

Acetylcholine, the transmitter liberated at parasympathetic nerve endings (Chapter 15), is so efficiently destroyed by cholinesterase that the amount in the general circulation is negligible. When injected in large doses it causes transient arteriolar vaso-dilatation. Vaso-dilator substances such as **histamine** and **adenosine triphosphate** (ATP) occur in the tissues, but only in insignificant amounts in the general circulation. Large quantities of **post-pituitary** extracts (as compared with any likely to be released from the gland) produce in man a slight rise of blood pressure and intense pallor of the skin, due to constriction of the capillaries. There is no evidence that the posterior pituitary takes part in maintaining capillary tone in normal conditions (see Chapter 11).

The Circulation to Particular Organs

Methods of Determining the Blood-flow through an Organ. In *animal experiments*, one of several methods may be used :—

(1) A cannula is tied into the vein draining the organ ; the blood is collected for a given period of time and its volume measured.

(2) An instrument for measuring the rate of the blood-flow is interposed between the cut ends of the artery or vein supplying the organ ; for example, a rotameter or an electromagnetic flowmeter. The rotameter is essentially a slightly funnel-shaped vertical glass tube containing a float. The height to which the float is lifted by the blood stream is an index of the rate of flow. To express it is absolute units (say ml./min.), the instrument must be calibrated. The electromagnetic flowmeter works on the principle that an electromotive force is induced in any conductor (blood for example) which is moving in a magnetic field. The blood passes through a hole in a perspex block which is pierced on each side by the poles of a powerful magnet, and above and below by a pair of electrodes. The E.M.F. developed between the two electrodes increases in direct proportion to the velocity of flow. After amplification, it is recorded by an ink-writing pen-recorder or by a cathode ray oscilloscope and camera. This instrument, also, must be calibrated, but it has the advantage of responding accurately to very rapid changes of flow.

Two other methods, while they do not actually measure the rate of blood flow, may be used to detect changes in flow, or calibre of the blood vessels :—

(3) The organ may be placed in an airtight box for recording changes in its volume (plethysmograph). These represent changes in the volume of blood in the organ, and may be ascribed to active changes in the calibre of its vessels so long as passive changes due to alterations of the general arterial and venous pressures can be excluded.

(4) In a transparent tissue lying superficially, like the frog's web, or the conjunctiva, or in an organ that can be exposed, the vessels may be observed microscopically and their diameters measured.

In man, extensive operative interference is not possible, and one of the following less direct methods must be used :—

(1) The generalised Fick principle, as already described (p. 23 above). The rate of blood flow through the organ considered is given by $F = Q(c_a - c_v)$, where Q, c_a and c_v have the same meanings as before. Blood from any artery can be used for the estimation of c_a; for that of c_v the blood must come from the venous outflow of the organ under consideration. This principle is used for estimating the rates of blood flow through the human hepatic, renal, cerebral and coronary circulations. If the reference substance is dissolved in the plasma only, the value of F obtained is that of the rate of *plasma* flow: the total blood-flow is obtained by dividing this by the relative volume of the plasma, which may be obtained by means of the hæmatocrit (Chapter 3, p. 70).

The human *hepatic blood-flow* is estimated by using the dye bromsulphalein as reference substance: this is excreted by the liver into the bile. The value of Q is given by the rate of intravenous infusion; c_a is obtained from an arterial blood sample; and c_v from a sample obtained by means of an X-ray opaque non-wettable catheter introduced into an elbow vein and manipulated

TABLE 2. 3

Approximate Distribution of a Cardiac Output of 5·0 litres per minute in a Man at Rest

	Weight kg.	Blood Flow (ml. per min.)	
		Total	Per 100 g, Tissue
Brain	1·5	750	50
Heart	0·3	150	50
Liver	1·5	1,500	100
Kidneys (2) . . .	0·3	1,200	400
Skeletal muscles . .	25·0	750	3
Other organs . . .	40·0	650	1·5

into the openings of one of the hepatic veins. The results, expressed in ml./min., are given in Table 2. 3. Para-aminohippuric acid (PAH), or alternatively diodone, is used for estimating the human *renal blood-flow*. The value of Q is obtained by estimating the amount excreted in a sample of urine collected during a known time. The value of c_v may be obtained from a sample of renal venous blood obtained by means of a catheter manipulated into one of the renal veins: but this is not usually done, since it is found that if the arterial concentration is not too high, the value of c_v is so small as to be negligible. PAH and diodone are removed from the plasma flowing through the kidney almost to completion: this will be referred to again in Chapter 9. Table 2. 3 shows the normal rate of renal blood-flow. To estimate the *rate of the cerebral circulation*, the subject is made to breathe air containing a small proportion (about 15 per cent.) of nitrous oxide. The nitrous oxide accumulates in the brain tissues, and after about ten minutes a state of equilibrium is reached. The total quantity of nitrous oxide taken up by unit mass of brain is then given by the product of the solubility coefficient (discovered in separate, *in vitro*, experiments, or on experimental animals), and the partial pressure of nitrous oxide in the venous blood at the end of, say, the tenth

minute. This quantity is equal to $10 \times Q/W$, if the determination has lasted exactly ten minutes, where W is the weight of the brain (not, of course, accurately known). The determinations of the arterio-venous concentration difference $(c_a - c_v)$ is more complicated, since initially, when there is no nitrous oxide in the blood, it is zero, and finally, when equilibrium has been reached, with the tissues, it is again zero. It is necessary, therefore, to find the effective average value during the intervening period. This is best done by plotting the values of c_a and c_v, determined every few minutes, against time, as in Fig. 2. 13. Any number of values of $(c_a - v_v)$ may be read from the smoothed curves drawn through these points and the arithmetic mean calculated ; but in practice, it is sufficient to take the values at the end of each minute. We can now apply the equation of the Fick Principle, but since we

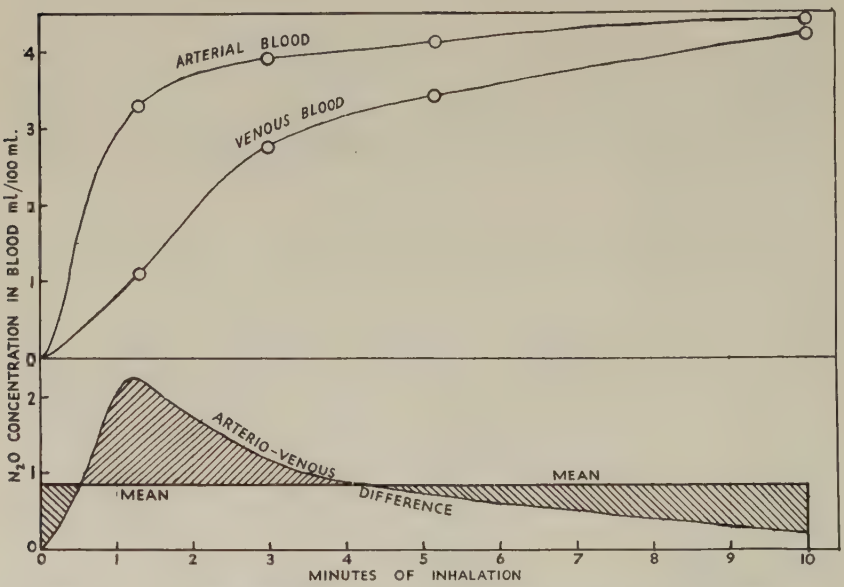


Fig. 2. 13. Typical Curves showing the **Concentrations of Nitrous Oxide** in the **Arterial and Jugular Venous Bloods** of a man during a ten-minute period of inhalation of 15 per cent. nitrous oxide. From curves such as these, the mean value of the arterio-venous concentration difference can be calculated, as is indicated in the lower part of the diagram. (After Kety and Schmidt.)

can only discover Q/W, and not Q, we can only measure F/W—*i.e.*, the rate of blood-flow through unit weight (usually 100 g.) of brain tissue. The cerebral venous samples are obtained from a needle placed in the jugular bulb just below the exit of the internal jugular vein from the skull. The insertion of the needle is done under local anæsthesia, and is safe in expert hands. Table 2. 3 shows the normal result. The nitrous oxide method has been used occasionally to estimate the human *coronary flow*. The venous blood samples are obtained from a catheter in the coronary sinus.

(2) The Venous Occlusion Plethysmograph. Part of an extremity is enclosed in a plethysmograph (Fig. 2. 14) which is a rigid watertight case. Any change in the volume of blood in the part enclosed is transmitted to a sensitive volume recorder. The blood-flow is measured by

recording the rate of increase of volume during temporary occlusion of the venous drainage. This is done by throwing a pressure of about 60 mm. Hg into a pneumatic cuff surrounding the limb just above the plethysmograph.

The following method may be used to detect changes in blood-flow :—

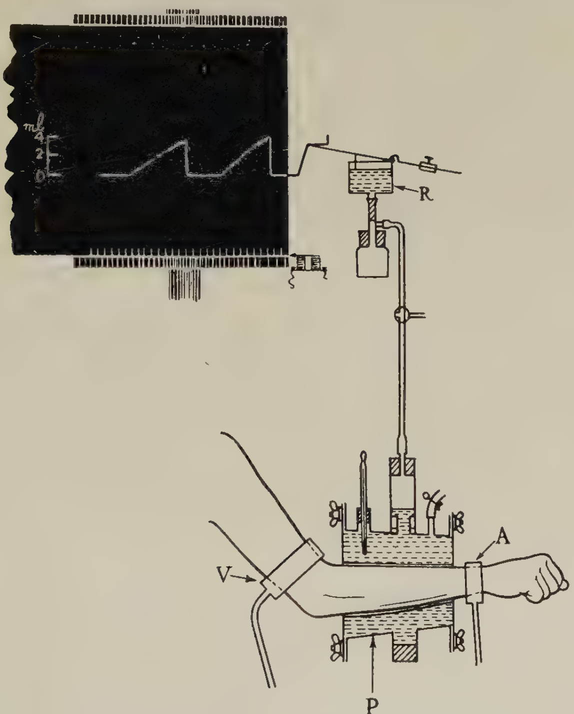


Fig. 2. 14. Determination of the **Rate of Blood-flow** in the **Human** Forearm by **Venous Occlusion Plethysmography.**

The plethysmograph (P) is filled with water maintained at a temperature of 34–35° C. To record the blood-flow, a pressure of 200 mm. Hg is thrown into the lower cuff (A) to arrest the circulation in the hand. A minute later a pressure of 60 mm. Hg is thrown into the venous occlusion cuff (V). The arterial inflow is recorded on the kymograph. After the experiment, the recorder (R) is calibrated, the speed of the paper is ascertained, and the forearm volume found by water displacement. From the slope of the inflow tracing, the rate of the blood-flow is calculated and expressed in ml. per minute for each 100 ml. of forearm. (From Starling's "Principles of Human Physiology.")

(3) The finger or hand or foot may be placed in stirred water in a calorimeter : the rate of heat elimination is proportional to the rate of blood flow. (A record from such a measurement will be found in Fig. 20. 3.)

The Pulmonary Circulation. The whole output of the right ventricle is delivered into the pulmonary artery at a pressure of 15 to 20 mm. Hg. The blood-flow through the lungs is the same as the outflow through the

aorta, *i.e.*, at rest about 4 litres per minute ; in exercise it may rise to 30 litres per minute or more. The resistance offered by the pulmonary vessels appears to be low. Since there is no alternative route for the blood between the right and left sides of the heart, it is not surprising that the vaso-motor supply is unimportant, though it appears that the vagus can dilate and the sympathetic constrict the vessels of the lungs.

The chief factor modifying the pulmonary circulation is the pressure change accompanying respiration. The lungs are elastic structures which are kept open by the chest wall ; if the chest is opened the lungs collapse. In consequence of this pull of the lungs, if a needle is thrust into the pleural cavity of man and connected to a manometer, this will normally register a pressure about 4 cm. H_2O below that of the atmosphere. During inspiration the lungs are further stretched and the intrapleural pressure falls by as much as 20 cm. water. These negative pressures affect all intra-thoracic structures but particularly the pulmonary capillaries, the great veins and the chambers of the heart in diastole, whose walls are yielding ; the effect on the thick-walled arteries is of less consequence. The effects of respiration on the systemic arterial blood pressure are complex and variable. In man during quiet respiration of the thoracic type (*i.e.*, mainly by the intercostal muscles) the blood pressure falls during inspiration and rises during expiration. This effect which is usual in man is probably due to mechanical changes in the pulmonary vessels ; when the lungs expand the pulmonary vessels are pulled open and fill with blood, and diversion of this extra blood from its onward movement leads to a reduced flow into, and output from, the left ventricle. Occasionally in man when the breathing is purely abdominal in type (*i.e.*, mainly by the diaphragm), the blood pressure rises during inspiration and falls during expiration. This effect may possibly be explained by the flow of blood into the thorax. During inspiration blood is aspirated into the great veins and heart from extra-thoracic structures, and the rise of intra-abdominal pressure produced by descent of the diaphragm forces blood from the abdomen into the chest ; in this way the filling and output of the heart would be increased during inspiration. X-ray photography, after intravenous injection of an opaque substance, shows that during inspiration the flow through the superior vena cava is increased, but in the dog, cat and rabbit the inferior vena cava is constricted by contraction of the diaphragm ; in these animals at least it is unlikely, therefore, that the flow through the inferior vena cava is increased during inspiration.

The pulse is accelerated during inspiration and slowed during expiration—a reflex effect which is abolished by section of the vagi.

The Brain. While the cerebral vessels are to some extent influenced by vaso-dilator metabolites during cerebral activity, they are little affected by vaso-motor impulses ; persistent inquiry failed to reveal any vaso-motor supply, until it was shown that stimulation of the cat's cervical sympathetic produced a slight narrowing of the pial arteries observed through a glass window screwed into the skull (Fig. 2. 6, p. 44) The cerebral blood-flow is thus determined in the main by the height of

the arterial blood pressure, and it is rather surprising that this should be regulated exclusively by receptors lying outside the brain, in the corotid sinus and arch of the aorta. For if these receptors are excluded, then alterations in blood-flow to the brain lead to no reflex changes altering the height of the arterial pressure, unless the cerebral blood-flow is so reduced that it produces asphyxia of the vaso-motor centre. Table 2. 3 (p. 57) gives some data about the cerebral circulation in man.

The importance and significance of the reflexes controlling blood pressure is now evident, for the brain is the master organ of the body and is extremely sensitive to reduction of its blood supply ; if the blood-flow to the brain ceases for five seconds consciousness is lost and after twenty seconds epileptic twitching begins. By means of the carotid sinus and depressor reflexes the arterial blood pressure, and thus the cerebral blood-flow are maintained by appropriate regulation of the rate and force of the heart-beat and of the blood-flow through organs other than the brain.

The Coronary Circulation. The heart muscle in mammals is supplied with blood from two coronary arteries arising from the aorta just beyond the semilunar (aortic) valves. The blood is returned to the right auricle by a number of openings of which by far the largest is the coronary sinus. The rate of the coronary blood-flow in an experimental animal may be measured by inserting a cannula into the coronary sinus ; the blood issuing represents, in the dog, three-fifths of the total flow through the whole coronary system.

The rate of flow in the coronary circulation has been measured in the human subject at rest (p. 58 above), and is about 150 ml. per min. It is not possible to measure it during severe exercise ; but to supply the heart with enough oxygen to produce the energy needed for a cardiac output of 30 litres per minute, the coronary flow would have to be nearly 1 litre per minute. Since the arterial blood pressure in severe exercise increases from 120/70 only to, say, 180/70, the six-fold increase in coronary flow cannot be due to an increased perfusion pressure. It is probably due to the vaso-dilator action of metabolites produced by the heart muscle during a condition of lowered oxygen pressure. The effect of lowered oxygen pressure is illustrated in Fig. 2. 15. When, in the heart-lung preparation, the lungs were ventilated for a short time with nitrogen, the coronary blood-flow increased about 10 times. The condition is not quite the same in exercise because the arterial blood is almost fully saturated with oxygen. Nevertheless the enormous increase in rate of oxygen usage by the heart must, at least temporarily, lower the oxygen saturation of the venous blood and the oxygen pressure of the tissue fluid bathing the plain muscle of the arterioles. This may relax them directly ; or they may be relaxed by vaso-dilator metabolites diffusing in from the heart muscle.

The autonomic nervous control of the coronary vessels is probably weak and its action easily swamped by the effects of local deficiency of oxygen. Stimulation of the sympathetic causes vaso-dilatation, and

stimulation of the vagus causes vaso-constriction ; but these effects may be due to concomitant alterations in heart metabolism. In the fibrillating heart, which no longer beats, the opposite effects are obtained during nerve stimulation. Probably, therefore, the sympathetic has a constrictor action, and the vagus a dilator one.

For a short time at the beginning of systole, contraction of the heart muscle arrests the coronary circulation. Studies of the rate of inflow into one of the coronary arteries show that the flow is intermittent—most rapid in diastole, and stopped during the isometric contraction

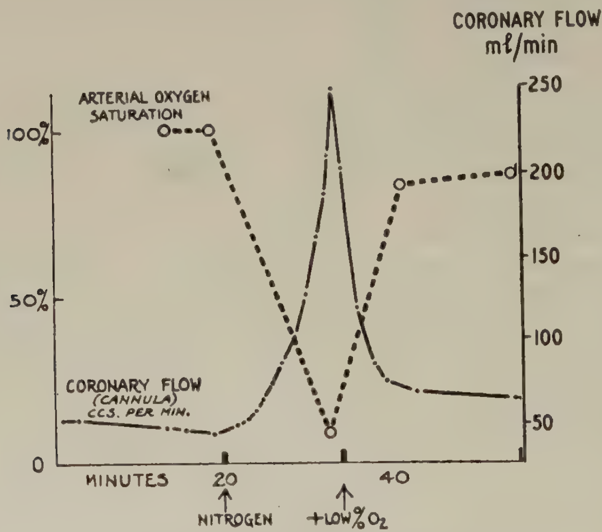


FIG. 2. 15. The Relation between **Coronary Blood Flow** and **Oxygen Saturation** of the Arterial Blood.

Observations on a heart-lung preparation. Between the first and second arrows the lungs were ventilated with pure nitrogen, and from the second arrow onwards with nitrogen to which a little air had been added. The coronary flow varies inversely with the oxygen saturation of the blood. (Hilton and Eichholtz.)

phase of the cardiac cycle, the coronary branches being compressed and occluded between the tightly contracting muscle strands. In this respect, cardiac muscle resembles skeletal muscle, in which, too, the flow is arrested during strong contraction.

Coronary Occlusion and Angina Pectoris. The function of the heart, like that of every organ, is intimately dependent on its blood supply and the arrangements that we have discussed are such that in health, increased work of the heart is accompanied by increased blood-flow. If a coronary artery is suddenly blocked by a clot, then the patient may die at once from ventricular fibrillation or after some hours from congestive heart failure ; in a large number of cases, particularly if the area deprived of its blood supply is small, the remaining healthy heart muscle is adequate to maintain the circulation at rest, the bloodless area is slowly converted into fibrous tissue and the patient recovers. Such small coronary occlusions are accompanied by intense substernal pain, probably due to the stimulation of sensory nerves

in the heart itself by chemical substances released locally from the ischæmic muscle. If the coronary arteries are thickened and narrowed by disease, they are incapable of dilatation, and the circulation becomes inadequate to the demands of muscular work. In this condition substernal pain (angina pectoris) is produced on exercise, probably again by the release of metabolites from the inadequately oxygenated heart muscle. A somewhat similar pain known as intermittent claudication is experienced in the muscles of the legs on walking, when the arteries are narrowed or blocked by disease. The student may reproduce this pain by working the muscles of his forearm, after the circulation has been arrested by inflating a cuff on the upper arm to above systolic pressure. He may ascertain that the rate at which pain develops depends on the frequency and the force of the muscular contractions. On stopping work the pain remains present until the circulation is restored, when it quickly disappears. The pain thus seems to be due to stimulation of the nerve endings in the muscles by a substance released during muscular contraction, and normally removed by the circulating blood.

The Skeletal Muscles. The changes in muscle blood-flow during exercise are of particular interest and are the outcome of two opposing factors—vaso-dilatation and mechanical compression of the vessels. In weak sustained contractions, the blood-flow increases ; but when the contraction is strong, as is that of the human gastrocnemius and soleus muscles, for example, when a person is standing tiptoe on one leg, the circulation in the muscle is almost arrested by the pressure of the taut tissue. It is not surprising that such contractions can only be kept up for three or four minutes. In rhythmic exercise, the muscle vessels dilate, but the pattern of flow depends upon the kind of movement. Blood-flow is continuous if the contractions are weak, but intermittent if they are strong.

During exercise, muscle blood-flow may rise ten-fold (Fig. 2. 16). The multilayered arterioles, the site of the main resistance (see Fig. 2. 4, p. 40) are widely dilated and the number of open capillaries increases many times, as do their diameters. Capillary surface area in the human gastrocnemius and soleus muscles may increase from the size of a handkerchief to that of a sheet.

As Fig. 2. 16 shows, this vaso-dilatation occurs in sympathectomised subjects, and is due to a local factor. In exercise, it is probable that the greatly increased rate of oxygen usage leads to a reduction in the oxygen pressure in the neighbourhood of the arterioles. Like the coronary vessels, the muscle vessels will be relaxed, either directly by the lack of oxygen or by vaso-dilator metabolites from the muscle fibres, perhaps for example, by adenosine triphosphate (ATP) or adenosine diphosphate (ADP). It is unlikely that alteration in the local acidity (pH), or in the local carbon dioxide or lactate concentration, is adequate to account for the vaso-dilatation. Thus the dominant feature of the control of the vessels supplying skeletal muscles, like that of the vessels supplying the heart, is a local control of the circulation in accordance with metabolic requirements.

The vaso-motor centre and sympathetic vaso-constrictor nerves maintain in resting skeletal muscle a small amount of constrictor tone. A change from the recumbent to the upright position brings reflexes

into play to prevent fall in arterial blood pressure, and the vaso-motor centre constricts the vessels in skeletal muscles. Reflex vaso-constriction also occurs during hæmorrhage. In the common faint (vaso-vagal syndrome) the fall in arterial blood pressure is due to vaso-dilatation in the skeletal muscles, mediated by the sympathetic, and unaccompanied by compensating vaso-constriction elsewhere. In exercise, the effect of the vaso-motor centre on skeletal muscle vessels is of little significance and easily swamped by the local vaso-dilator mechanism. The importance of this mechanism is shown, for example, by some observations on a policeman, aged 23, on whom bilateral lumbar sympathectomy

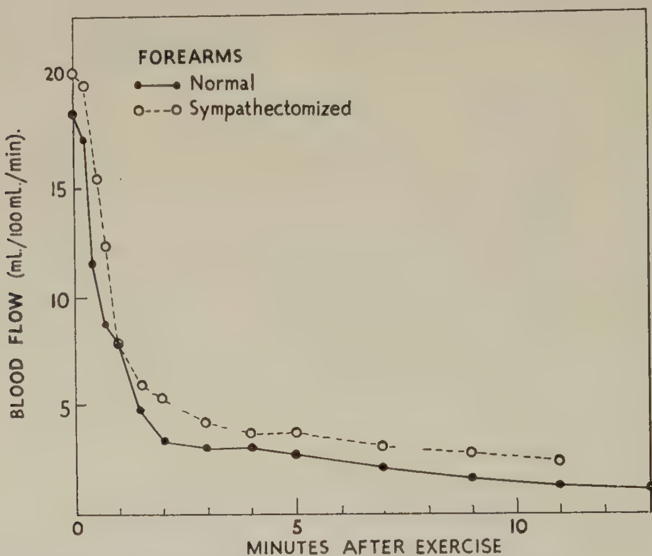


FIG. 2. 16. Results showing that exercise causes vasodilatation in human muscles ; the vasodilatation occurs in sympathectomised muscle and is due to the action of vasodilator metabolites. Venous occlusion plethysmography. The exercise was gripping a bar tightly in each hand for forty seconds. (From data by Grant, 1938.)

was performed in order to prevent excessive sweating of the feet. On the day before the operation he twice ran 380 yards in 65 and 61 seconds, respectively ; on the 99th day after the operation, he twice ran the same distance in 60 and 62½ seconds, respectively.

When the amount of adrenaline in the general circulation increases, the muscle vessels dilate : but they constrict when adrenaline is applied to them directly. That is to say, intravenous and intra-arterial infusions of adrenaline in man have opposite effects on the muscle circulation. The reason is not yet known.

The Salivary Glands. Stimulation of the parasympathetic nerve to the submandibular gland causes secretion and marked vaso-dilatation. Activity of the salivary secreting cells is accompanied by the release from them of a proteolytic enzyme : and in the presence of this enzyme,

tissue fluid protein is hydrolysed, forming a polypeptide known as *bradykinin*. This substance is a potent vaso-dilator, and its action on the neighbouring vessels increases submandibular blood flow in accordance with the metabolic requirement. It is probable that local vaso-dilator mechanisms are brought into action during the secretion of all other digestive glands, but this is not known for certain.

The Liver and Portal System. The blood-flow through the liver is very large (Table 2. 3, p. 57); about half the blood-flow through the inferior vena cava comes from this source. After leaving the intestinal capillaries the blood gathered into the portal vein traverses a second set of capillaries in the liver. Since the portal pressure in the dog is only about 8 cm. H_2O the resistance offered by the liver vessels must be very small. The liver is also supplied by the hepatic artery, which contributes about a quarter of the blood and 40 per cent. of the oxygen supplied to the organ. The liver, the portal vein and the territory it drains ordinarily accommodate about one-third, or more, of the total blood volume. Experiment suggests that a large proportion of this is expelled in the early stages of hæmorrhage. After the injection of adrenaline or stimulation of the sympathetic nerves the outflow from the liver exceeds its inflow, a large part of its blood being thus discharged into the great veins. It is likely, therefore, that the liver and portal system constitute a variable reservoir, whence blood is discharged to augment the inflow and output of the heart in conditions such as hæmorrhage, emotion and exercise.

A piece of dog's colon, transferred to the outer abdominal wall with its nerve and blood supply intact, blanches at the beginning of exercise, though, as exercise is continued, it slowly fills again with blood. This presumably illustrates what happens to the vessels of the whole gut in exercise.

The Spleen. The branches of the splenic artery open into venous sinuses, which unite to form the splenic vein. Along the course of the artery and vein are perforations communicating with the spleen pulp, which contains red and white blood corpuscles in its network. In the dog the spleen has been brought to the exterior through an incision on the abdominal wall and left there for many months. Its size at rest indicates that it may hold one-fifth of the total blood volume. During hæmorrhage, emotion, asphyxia and muscular exercise the muscular capsule of the spleen contracts, and its blood content, which is exceptionally rich in red cells, is expelled into the general circulation. The reservoir function of the spleen appears to be less important in man.

The Kidneys. The regulation of the renal circulation is described in Chapter 9. Some data about its normal rate in man are given in Table 2. 3, p. 57.

The Skin. Heat is lost from the skin by conduction, radiation and evaporation. The rate of heat loss is regulated by the temperature regulating centre acting through the sympathetic nerves to the cutaneous blood vessels and sweat glands. (This will be discussed more fully in Chapter 20.) According to the manner of this regulation, the skin of the body can be divided into two areas. Thus the skin of the hands,

feet, nose, lips and ears has to be considered separately from that of the forehead, trunk and limbs as far as the wrists and ankles.

When the room temperature is below about 20° C. (68° F.), the vessels in the hands, feet, lips and ears of an ordinarily clothed man are constricted—strongly so if the temperature is below about 16° C. (about 60° F.). The effect of a rise of environmental temperature on the blood-flow through the skin of the hand is shown in Fig. 2. 17. This flushing of the hand is chiefly due to a release of the vaso-constrictor tone brought about by sympathetic nerves, because blocking the nerve

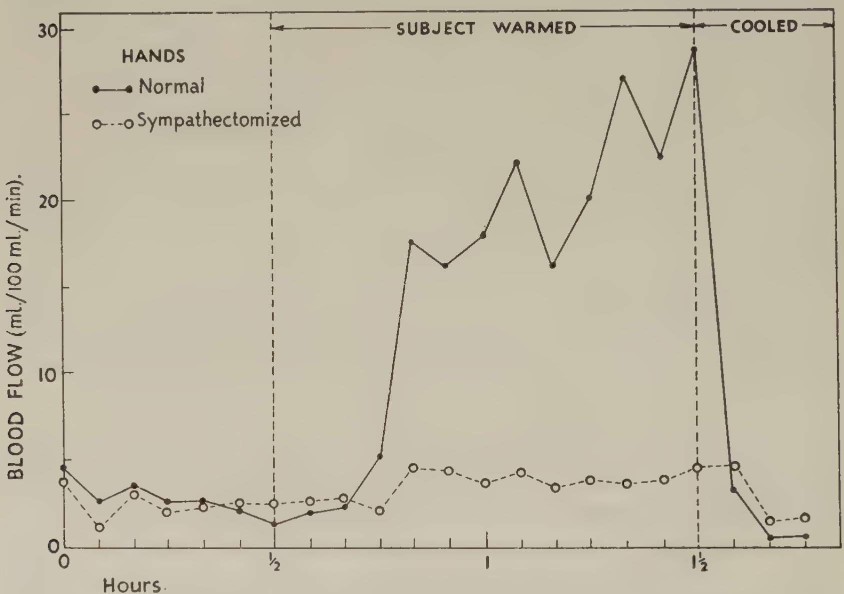


FIG. 2. 17. Results showing that **Warming the body causes Vasodilatation** in the hand ; this does not occur in the sympathectomised hand. It is mainly, if not entirely, due to the release from sympathetic constrictor tone. Venous occlusion plethysmography. The body was warmed by covering the subject with blankets and immersing the feet in water at 44° C. ; and cooled by removing the blankets and immersing the feet in water at 17° C. (Barcroft.)

supply by local anæsthesia causes a similar increase in flow. The chief factor in the regulation of the skin blood-flow in accordance with changes in the environmental temperature is a central mechanism which responds to a rise in blood temperature by reduction of sympathetic tone, and *vice versa*, as discussed in Chapter 20. In this connection an important part is played by the arterio-venous communications, short wide connections between the arterioles and venules, which are very abundant in man in the nail beds, the skin covering the volar surfaces of the fingers and palm of the hand and in the corresponding sites of the foot. The changes in blood-flow through the fingers which may occur in response to changes in body temperature are very large,

ranging from 1 ml. per minute through each 100 ml. of finger when the body is cool, to 100 ml. per minute when the body is hot.

The skin of the forehead, trunk and proximal parts of the limbs has very little, if any, sympathetic vaso-motor innervation. Nevertheless, in a very warm subject marked vaso-dilatation occurs, though without much flushing because the vessels lie too deeply. This vaso-dilatation accompanies sweating. It is due to the vaso-dilator action of bradykinin, formed during sweat gland activity. It will be recalled that the vaso-dilatation which accompanies salivary secretion is also due to the formation of bradykinin (p. 65).

The colour and temperature of the hand and face are closely related to the state of the skin circulation. The hot pale hand, common in summer, is one in which blood flows rapidly through the deeply situated invisible arterio-venous anastomoses in the fingers and palms. The superficial capillaries whose contents give colour to the skin are narrow and poorly filled. The cold red hand seen in winter is one in which the arterio-venous anastomoses and arterioles are narrow and the capillaries dilated. The blood remains red, since little oxygen is removed by the cold tissues. The cold blue hand is one in which the arterial vessels are still further constricted, and the flow becomes so slow that an appreciable fraction of the oxygen content of the blood is removed even by the slowly metabolising tissues.

Although the vaso-motor and temperature regulating centres predominate in the regulation of the skin circulation—directly through vaso-motor nerves or indirectly through local vaso-dilator substances—there are some other responses which must be briefly described.

(1) *Reactive Hyperæmia.* This important response was first seen in the arm and leg. If the circulation to a warm limb is arrested for a few minutes and then released, a bright flush, reactive hyperæmia, at once suffuses the skin and then slowly fades. After circulatory arrest lasting ten minutes the blood-flow to the forearm may be increased to ten or twenty times the normal ; both muscle and skin vessels share in this vaso-dilatation. The intensity and duration of reactive hyperæmia depend on the duration of circulatory arrest and on the temperature of the limb. The flush represents a dilatation of the minute vessels ; it is independent of any central or local nervous mechanism and is due to the action of vaso-dilator substances formed locally and normally removed by the circulating blood. When, as is constantly happening, areas of skin and of subcutaneous tissue are rendered bloodless by supporting the weight of the body, they may be said to accumulate a blood-flow debt ; reactive hyperæmia ensures the discharge of this debt as soon as blood is free to enter the tissues again.

(2) *The White Reaction.* This is chiefly of interest as the basis of an experiment performed by Lewis to show that human capillaries can contract actively. If the skin of the forearm or back is lightly stroked with the end of a ruler, the line of the stroke becomes marked by pallor. This is due to narrowing of the minute vessels (capillaries and venules), for these are the only vessels which come near enough to the surface

of the skin for blood within them to be visible. Narrowing of the capillaries in this white reaction might be due to (*a*) constriction of the deeper arterioles and passive collapse of the more distal capillaries ; or to (*b*) active contraction of the capillary walls. Lewis showed that the white reaction can be induced in the forearm skin after the circulation in the arm has been arrested. In this case, the pallor cannot be due to constriction of the underlying arterioles, for this would, if anything, tend to increase capillary pressure ; therefore the capillary walls themselves must contract.

(3) *Triple Response.* If the skin is injured by scratching, by lightly burning, by freezing or by pricking in injurious substances such as hydrochloric acid or caustic soda, the point of injury is marked by reddening of the skin, which later gives place to whealing as fluid passes out of the capillaries and distends the tissue spaces of the skin. Around the local reddening is a diffuse bright red mottled flush or " flare," which

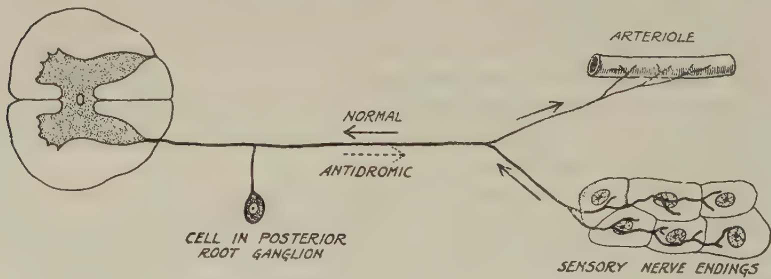


Fig. 2. 18. Diagram of the Nervous Connections concerned in the **Axon Reflex.**

is due to the opening of the surrounding arterioles. The local redness due to widening of the minute vessels (capillaries and venules), the wheal due to their increased permeability, and the flare are the components of the triple response of the vessels to injury. The whole response is independent of the central nervous system, being unchanged immediately after section of all nerves to the skin. After all the nerves have degenerated, however, the flare is absent ; if the sympathetic supply alone has degenerated the flare is unimpaired. The flare is an example of a local axon reflex through the sensory fibres. The fibres entering the posterior roots of the cord divide at their periphery into branches to the blood vessels and to the tissues. Injury to the skin stimulates the sensory branches, and the stimulus passes proximally to the point of bifurcation and back down the other branch to the arterioles (see Fig. 2. 18.) The whole of the triple response has been shown to be due to the release of a chemical substance from the injured skin, and this, from its resemblance to histamine, has been termed " H substance." The triple response is, in the skin, the vascular basis of the phenomenon of inflammation.

CHAPTER 3

CARRIAGE OF OXYGEN AND CARBON DIOXIDE

THE blood, in its circulation, brings to the tissues the foodstuffs and oxygen necessary for their metabolism, and carries away the carbon dioxide and other waste products. The blood, however, is no mere indifferent circulating fluid, but has remarkable " buffering " properties ; these enable it to change its composition with respect to the most important substances—oxygen and carbon dioxide—without at the same time bringing about an equally large change in the composition of the fluids surrounding the tissue cells. It can transport much larger quantities of these substances than could a simple saline solution, for example, circulating at the same rate. This property, which is clearly of great importance to the " efficiency " of the blood as a carrier of oxygen and carbon dioxide, is due to the presence of the red " respiratory pigment " **hæmoglobin.** Indeed, in the absence of a substance which has the remarkable physical, chemical and physiological properties of hæmoglobin, the only animals which could exist would be small and sluggish, living in well-aerated water. One other substance, only, is known which has these properties, and that is *hæmocyanin*, the blue " respiratory pigment " of the molluscs and some arthropods.

The Red Blood Corpuscles

When observed under the microscope, blood is seen to consist of an enormous number of pale yellow discs, which are the red blood corpuscles (also called **erythrocytes**, or **red cells**), floating in the clear colourless plasma. (When seen in bulk, in much thicker layers, the plasma is yellow, and the erythrocytes red.) The cells are described as biconcave discs, since they are thicker near the edge than in the middle ; their average diameter in human blood is about $8{\cdot}6\mu$ and their thickness about $2{\cdot}6\mu$. Their specific gravity is greater than that of the plasma (about 1·10 against about 1·03), which accounts for the fact that they settle out when blood is allowed to stand, or is spun in a centrifuge. The hæmoglobin, which is responsible for the red colour of blood, is thus contained entirely within the red blood corpuscles. This does not affect appreciably its properties as an efficient carrier of oxygen and carbon dioxide : but if it were in simple (colloidal) solution, it would be lost from the blood through the kidneys (Chapter 9). In some circumstances, the viscosity of a suspension of red cells is smaller than that of a solution containing the same quantity of hæmoglobin in unit volume (Chapter 2, p. 33), so that less power is needed to drive it round the circulation at the required rate.

The estimation of the number of red cells per cubic millimetre of blood (the *Red Cell Count*) is important for both clinical and research

purposes. The apparatus used for this purpose is called a *hæmocytometer* (from the Greek for " blood-cell measurer ").

The blood is first diluted 200-fold in a special pipette with " Hayem's fluid " which " fixes " the corpuscles (in the histological sense). A special microscope slide is then used, in the middle of which is a polished surface lying exactly 1/10th mm. below the general level of the slide. A drop of the diluted blood is placed on the surface, and a cover slip lowered over it, so that it spreads and forms a film exactly 1/10th mm. thick. On the floor of the slide are engraved two sets of fine lines, 1/20th mm. apart, crossing at right angles to form a large number of squares. When the red cells have settled down, they are counted under the microscope, those in each square having previously been suspended in $(1/20 \times 1/20 \times 1/10) = 1/4{,}000$ cu. mm. of fluid. To obtain the average number of cells per cubic millimetre, in the original blood, about 500 cells are counted ; the count is first divided by the number of squares that the cells occupy, then by the volume of fluid above each square (1/4,000 cu. mm.), and is finally multiplied by the dilution factor (200).

The red cell count in healthy men varies from $4\frac{1}{2}$ to 6 million per cu. mm., the average being about $5\frac{1}{2}$ million. In women the count is even more variable, but on an average about 10 per cent. lower than in men. The uncertainty in the method of obtaining a blood count, as ordinarily performed, is likely to be at least $\pm$ 5 per cent.

The red blood corpuscles normally occupy 40 to 45 per cent. of the volume of the blood in man. This important ratio can be measured by means of the *hæmatocrit* (from the Greek for " blood-separator "). This is a graduated capillary tube of uniform bore, which is filled with blood and centrifuged at high speed until the column of sedimented red cells shows no further shrinkage ; the volume of the tightly packed cells can then be read off and compared with that of the whole blood.

The microscopic examination of the blood cells is conveniently carried out on dried and stained smears of blood. On drying, the red cells shrink, so that their diameters are diminished by 8 to 16 per cent. In such dried preparations of normal human blood the average diameter of the red cells is only $7{\cdot}4 \pm 0{\cdot}3\mu$. The shrinkage and distortion of the cells so prepared are of no great consequence in the clinical examination of the blood, because the characteristic changes in blood associated with different diseases are commonly described in terms of the changes which appear in the dried smears.

Red blood corpuscles have no nuclei, but are, nevertheless, living cells : they have quite active metabolic processes, and require energy for their existence, which they obtain, chiefly, by means of the anærobic breakdown of glucose. But they have a relatively short life-time, and old ones are continually being destroyed and replaced by new ones formed, in post-natal life, in the blood spaces of the bone marrow. In normal individuals, the rate of production is adjusted so as to be equal to the rate of destruction when the number of red cells in the circulating blood is such that the red cell count has its normal value of about 5 million per cu. mm. : but, following an abnormally large loss of blood for example after a hæmorrhage, the rate of production is greatly increased until the loss is made good. The primary stimulus for this is

the reduction in the rate of supply of oxygen to the tissue cells generally : with less hæmoglobin in each litre of blood, less oxygen can be carried round from the lungs. The production of red cells is accelerated, also, when the supply of oxygen to the tissues is deficient by virtue of an inadequate supply of oxygen to the lungs, as at high altitudes (this will be discussed further in Chapter 4). The bone marrow itself does not respond directly to the deficient supply of oxygen, but, apparently, to some substances in the circulating blood, produced elsewhere and not yet identified.

Many different methods have been used to estimate the average life-time of a red blood cell, and widely different values have been obtained. In principle, a certain proportion of the circulating red cells is " labelled " in some way, and the rate of their disappearance measured. (1) Cells may be injected intravenously which contain a different type of agglutinogen from that of the cells normally present (see Chapter 21), care being taken, of course, that the cells injected are not such as to be affected by agglutinins in the plasma of the recipient. (2) Cells which have been removed from the subject under investigation may be allowed to take up *in vitro* the radioactive isotope of phosphorus ^{31}P (as phosphate), or that of chromium ^{51}Cr (as chromate), and then returned to the circulation. (3) The hæmoglobin incorporated in the cells formed in the ordinary way in the bone marrow may be " labelled " by injecting intravenously, over a suitable period of time, either (*a*) salts of the radioactive isotope of iron ^{59}Fe, or (*b*) the amino-acid glycine in which the ordinary carbon or nitrogen atoms, ^{12}C and and ^{14}N, have been replaced by those of the isotopes ^{14}C and ^{15}N, as the case may be : the rate at which these cells appear may be measured as well as, or instead of, the rate at which they disappear. None of the various methods used is without its difficulties, and various corrections must be made to the results obtained. The most reliable measurements indicate an average life-time in man of some 100 to 130 days.

In certain abnormal conditions, the rate of production of the red cells may be unduly small, or the rate of destruction unduly large : in either event, in the steady state when the two rates are equal, the red cell count will be less than the normal value, producing *anæmia.*

(1) Study of the disease pernicious anæmia (and certain other related kinds of anæmia) has shown that for the normal development of the red cells two specific *vitamins* are necessary : these are known as *vitamin* B_{12} (which contains the rather uncommon element cobalt, and is also known as cobalamin) and *folic acid* (or pteroyl-glutamic acid). Both vitamins are present in adequate quantities in normal protein foods of animal origin, and particularly in liver ; folic acid is present also in fresh green vegetables, and vitamin B_{12} is synthesised by many kinds of micro-organism, including some of those in the alimentary tracts of mammals. Anæmia may result from malnutrition : but more usually, owing to some diseased condition of the alimentary tract, the vitamins are not properly absorbed and anæmia results even though the diet is fully adequate.

When given by subcutaneous injection, vitamin B_{12} is highly active in improving the condition of a patient with pernicious anæmia : but when given by mouth it is far less active. It would appear that it is not well absorbed from the alimentary canal unless first acted upon by a constituent

of the gastric juice, known as the *intrinsic factor* ; vitamin B_{12} being the *extrinsic factor*. This intrinsic factor is absent from persons suffering from pernicious anæmia.

(2) The cells of the bone marrow, like other actively dividing cells, are very susceptible to the action of *ionising radiations* and particles—X-rays and the γ-rays, α-particles, electrons and neutrons emitted during radioactive disintegration of unstable atomic nuclei. After excessive exposure to such radiations, cell division in the bone marrow ceases, or becomes disordered, the rate of production of normal red cells is greatly reduced, and abnormal kinds of cell may appear in the circulation.

(3) Iron is needed for the synthesis of the hæmoglobin within the red cells ; but the iron released from destroyed cells is used again in the production of new ones, and ordinary diets contain sufficient to make up for any deficiency. Small amounts of copper are essential to the proper utilisation of the iron. Anæmia, however, may occur in infants during the suckling period, owing to an almost complete lack of iron in the milk ; it can be cured by administration of salts of iron.

(4) The rate of destruction of the red cells may be abnormally great owing to a congenital defect in their structure or metabolic processes, which shortens their life-time ; or to the presence in the plasma, as a result of disease, of a specific " anti-body " which destroys them.

In the earlier stages of their development, red cells possess nuclei. Two stages in the development of nucleated red cells from the mesenchymal parent cell are distinguished ; they are known successively as *erythroblasts* and *normoblasts*. The normoblast becomes an adult red cell when it loses its nucleus. For the first few days of its adult life, spots and threads of basophilic material can be demonstrated in the red cell by vital staining with cresyl blue. In this stage it is known as a *reticulocyte* or reticulated red cell. After the first week of life reticulocytes form only about 1 per cent. of the circulating red cells, but they form a larger proportion of the red cells found in bone marrow. The proportion of them in circulating blood rises considerably whenever the new formation of red cells is increased, as after a hæmorrhage, or after the successful treatment of pernicious anæmia.

The hæmoglobin in the fragments of the old red cells is converted by the reticulo-endothelial cells into inorganic iron and the pigment *bilirubin*. These travel in the plasma to the liver, and the bilirubin, after a slight change, is secreted into the bile. In the bowel bilirubin is changed into *stercobilinogen* and *stercobilin*, which give the fæces their dark colour. Not all the bile pigment is excreted, however. Some is reabsorbed into the portal blood stream in the form of a colourless compound called *urobilinogen*, which is picked up by the liver, perhaps to be used in the production of new hæmoglobin. When large quantities of bilirubin are formed (as in hæmolytic jaundice), and consequently large amounts of urobilinogen absorbed, appreciable amounts of urobilinogen are excreted in the urine. (This is to be distinguished from the excretion of bilirubin in the urine which occurs when the bile passages are obstructed, for example, by a gall-stone (obstructive jaundice).) Urobilinogen is also found in the urine when the liver function is impaired, and thus unable to deal with the normal quantities carried to it in the portal blood stream.

Hæmoglobin

Chemically, hæmoglobin consists of an iron-porphyrin complex, known as reduced hæmatin, or *hæm*, united with a protein, *globin*.

Hæm is closely related chemically to the prosthetic groups of many enzymes concerned in the oxidation of the foodstuffs, notably the cytochromes, and peroxidase. Hæm can unite with a number of nitrogen-containing substances, forming *hæmochromogens* ; hæmoglobin is a special case of a hæmochromogen.

Hæm and the hæmochromogens contain iron in the *ferrous* (divalent) state. If treated with a suitable oxidising agent, the iron is oxidised to the *ferric* (trivalent) state, with the formation of *hæmatin* from hæm, and *methæmoglobin* from hæmoglobin.

Hæmatin unites with hydrochloric acid to form the hydrochloride *hæmin* ; this substance crystallises readily from solution, and the character of the crystals is sufficiently distinct to be used as a chemical test for blood pigments. Hæmoglobin itself only crystallises easily when derived from certain species, *e.g.*, the horse, rat or guinea-pig. The shape of the hæmoglobin crystals varies with the species of blood used. For further information as to the chemical relations of hæmoglobin, reference should be made to a text-book of bio-chemistry. All the compounds concerned have characteristic absorption spectra, which are invaluable for their identification.

Hæmoglobin combines with oxygen to form a scarlet compound, *oxyhæmoglobin* ; this contains 1 gram-molecule of oxygen for each gram-atom of iron in the hæmoglobin. The oxygen can be removed again, just as if it were in simple solution, by shaking the hæmoglobin solution repeatedly in a vacuum, or with gas containing no oxygen, such as hydrogen or nitrogen. The colour of the solution changes to purple when the oxygen is removed, and the hæmoblobin is then said to be reduced. Hæmoglobin, oxygen and oxyhæmoglobin are thus components of a reversible chemical reaction :

$$\mathrm{Hb} + \mathrm{O_2} \rightleftharpoons \mathrm{O_2Hb}$$

(Hb being the symbol generally used for hæmoglobin.) In this reaction, the oxygen is attached to the hæm part of the hæmoglobin molecule, but the iron atom remains in the bivalent state : the reaction, therefore, is not one of oxidation, in which the iron would become trivalent, but is referred to as *oxygenation*. Reactions of this kind, in which the oxygen combined with the hæm can be removed merely by removing the oxygen dissolved in the solution, occur only when the hæm is united to the particular protein, globin : other hæmochromogens may be oxidised, but not oxygenated, by oxygen in solution. It is the existence of this easily reversible reaction with oxygen that enables hæmoglobin to act as a carrier of oxygen. In the carriage of carbon dioxide, easily reversible reactions are again involved, but with the globin part of the molecule, rather than the hæm part ; they are not accompanied by colour changes, and so are less conspicuous. But it is one of the outstanding features in the properties of hæmoglobin that the hæm and the globin parts interact with one another. The globin not only allows the hæm to react reversibly with oxygen but small changes in its precise chemical composition, or state of ionisation, may affect considerably the affinity of the whole molecule for oxygen, as will be discussed later. Conversely,

the state of oxygenation of the hæm affects the ionisation of the globin, and its affinity for carbon dioxide.

For an adequate rate of transport of oxygen and carbon dioxide to and from the tissues and the lungs, it is essential that there should be an adequate concentration of hæmoglobin in the blood and that the hæmoglobin should be able to take up, and lose, oxygen and carbon dioxide. A man may become unduly distressed by going upstairs or running for a bus, not only because his heart or lungs are inadequate, but also because there is not enough hæmoglobin in his blood to carry the oxygen and carbon dioxide, or to prevent his tissues from becoming too acid. A blood count, of course, will indicate whether he has the normal number of red cells, but these may not contain the normal amount of hæmoglobin. The concentration of hæmoglobin in the blood may be measured in terms of the maximum quantity of oxygen with which unit volume (usually 100 ml.) will combine—*i.e.* the *oxygen capacity* of the blood, as described below (p. 76) : or, more simply, in terms of its colour, since hæmoglobin has so strong a red colour.

The colorimetric (or more strictly absorptiometric) estimation is usually made with the hæmoglobin in combination with carbon monoxide (carboxy-hæmoglobin—see p. 80 below) ; acid hæmatin and reduced hæmoglobin can also be used ; all three substances are more stable than oxyhæmo-globin. In the *Haldane Hæmoglobinometer*, the appropriate volume of the blood under test (20 cu. mm.) is treated with coal gas, so as to convert the hæmoglobin to carboxyhæmoglobin, and is then diluted in a special tube, of the proper size, until it has the same depth of colour as has an arbitrary standard ; this may either be a solution of carboxy-hæmoglobin in a sealed tube, or better, a solid rod of suitably coloured glass. The final volume of the diluted unknown blood is then read on special graduations on the tube, which give the hæmoglobin concentration as per cent. of that of the standard. This value, sometimes known as the " hæmoglobin concentration," does *not* give the actual hæmoglobin concentration of the blood, but only the relation to the standard. This latter is conventionally adjusted to contain 15·6 g.Hb per 100 ml. blood, which is equivalent to an oxygen capacity of 21 ml. O_2 per 100 ml. blood. These figures, therefore, will also apply to any blood which has " 100 per cent. hæmoglobin."

Alternatively, and more accurately, the hæmoglobin concentration may be measured directly in terms of the transmittance of the solution—*i.e.* the fraction of the incident light which is transmitted through unit depth of solution ; the intensities of the incident and transmitted lights being measured by means of a photo-electric cell. It is best to use light of such a colour that it is strongly absorbed by hæmoglobin. Solutions of oxyhæmo-globin absorb light in the whole of the blue-violet region of the spectrum (this is why they have a red colour) and in the extreme red end. There are also two well-defined absorption bands, one in the yellow and the other in the green, which are valuable for purposes of identification. Solutions of reduced hæmoglobin absorb light more strongly in the extreme red end of the spectrum than do solutions of oxyhæmoglobin, and less strongly in the blue-violet region, and there is only one band in the yellow-green region ; this is broad and dim. By the use of suitably coloured filters, transmitting light of certain wave-lengths only, and photo-electric cells for measuring the transmittance of the blood or hæmoglobin solution at these wave-lengths, it is possible to measure, independently, the concentration of oxyhæmoglobin and the con-centration of reduced hæmoglobin. The complete apparatus for doing this is known as an *oximeter* : it can be made small enough to be attached to the

lobe of man's ear, for example, and will give a continuous record of the degree of oxygenation of his blood.

The hæmoglobin content of the average red blood cell is expressed in terms of the *Colour Index.* The hæmoglobin concentration of the blood, as per cent. of the standard value, is divided by the red cell count, also as per cent. of the standard value (taken as 5 million per cu. mm., so that the per cent. red cell count is the actual red cell count multiplied by 20). The colour index is thus ordinarily about unity in healthy persons, but is decreased considerably in secondary anæmia (*e.g.* after repeated hæmorrhage), in which the hæmoglobin concentration may fall to 50 per cent of the standard value, and it is often increased in pernicious anæmia, in which the red cell count may fall to one million or less.

The Carriage of Oxygen

The amount of oxygen actually combined with the hæmoglobin in unit volume of some particular sample of blood—the *oxygen content*—is often expressed as a percentage of the amount that would be combined if it were fully saturated—*i.e.*, as the *percentage saturation.* It is often convenient, also, to refer to the degree of *unsaturation* of a sample of blood ; this is a measure of the extent to which the actual oxygen content falls below the oxygen capacity. Thus if, for example, the oxygen content were found to be, say, 10 ml./100 ml. blood, and the oxygen capacity of the same, or of an exactly similar, sample of blood were found to be, say, 15 ml./100 ml. blood, the percentage saturation would be 66·7 : the sample would have an unsaturation of 5 ml./100 ml. and the percentage unsaturation would be 33·3.

If hæmoglobin takes part in a reversible chemical reaction with oxygen, to form oxyhæmoglobin, it is to be expected that in any particular sample of blood, the percentage saturation would depend on the concentration of oxygen in solution in the blood (provided, of course, that the hæmoglobin is not fully saturated with oxygen). The concentration of oxygen in solution—the amount which is physically dissolved in unit volume—is directly proportional to the partial pressure of oxygen in the gas with which the blood is brought into equilibrium.

The partial pressure of a gas in a mixture of several gases is that fraction of the total pressure which can be considered as being contributed by that gas. This is considered in more detail in the next chapter (p. 108), where examples are given of the partial pressures of the gases in gas mixtures under various conditions.

It is thus usual to relate the quantity of oxygen combined with hæmoglobin to the partial pressure of oxygen in the gas with which the blood is equilibrated. This relation is shown in Fig. 3. 1 (p. 77 below). Curves such as these are known as *dissociation curves* (or, more properly, *equilibrium curves*). The points through which they are drawn are determined by shaking a sample of blood gently with a suitable mixture of oxygen, nitrogen and, if desired, carbon dioxide, until equilibrium is reached (thirty to forty-five minutes), then measuring the oxygen content of the blood and, by analysing the gas mixture, the partial pressure of oxygen. At the partial pressure at which the hæmoglobin is sensibly fully saturated (100 mm. Hg), the amount of

oxygen dissolved in the blood is about 1/70 of the amount combined with the hæmoglobin.

The type of *gas analysis apparatus* most generally used is that devised by Haldane. A sample of the gas to be analysed is drawn into a water-jacketed burette, previously completely filled with mercury, and its volume measured. Carbon dioxide is then absorbed by transferring the gas into a vessel from which it displaces a strong solution of caustic potash ; it is then brought back into the burette and its volume measured again. Oxygen is absorbed by transferring it in a similar manner into a vessel previously filled with an alkaline solution of pyrogallol, and the residual volume measured once more. Absorption of the carbon dioxide and oxygen is hastened by passing the gas to and from the appropriate absorbing vessels and the burette by raising and lowering the reservoir of mercury connected with the lower end of the burette. Various modifications have been made to the original design in attempts to make its use quicker and simpler. The Scholander apparatus, for example, is specially designed so as to need only 0·5 ml. of gas for analysis (the Haldane apparatus needs about 10 ml.). It has some advantages over the Haldane apparatus even when plenty of gas is available for analysis ; some find it easier to use and keep in proper order.

The *amount of oxygen combined with hæmoglobin* is estimated by driving off the oxygen from the blood, either : (1) by boiling *in vacuo* ; or (2) by adding ferricyanide ; or (3) by both together.

(1) The first was the method used by the earliest workers, who evacuated a large vessel by means of a mercury pump, for example, a Töpler pump, and ran a known volume of blood, previously warmed to 40°, into it. The gas liberated was then pumped off and delivered to a gas analysis apparatus, where its total volume and composition were determined.

(2) If oxy-hæmoglobin is treated with ferricyanide, the reduced hæmoglobin which is always present in small quantities is converted into methæmoglobin, so that the equilibrium between oxy-hæmoglobin and reduced hæmoglobin is upset ; in an attempt to restore equilibrium oxyhæmoglobin gives up its oxygen, and more reduced hæmoglobin is formed ; this is again removed by the ferricyanide, and more oxyhæmoglobin decomposes, and so on. In this way, the whole of the oxygen in reversible combination can be driven off under suitable conditions, and its volume measured. When this reaction was discovered, it seemed that it would be a much easier way of determining the oxygen combined with hæmoglobin than the vacuum pump methods previously used. Haldane, indeed, at once adopted the principle in his blood gas apparatus. In this oxyhæmoglobin is mixed with *alkaline* ferricyanide (the alkalinity prevents the escape from the solution of any carbon dioxide gas) and the only gas evolved is oxygen, which is measured directly in a burette over water. The same principle is also used in the Barcroft differential apparatus. Much work has been done by these two methods, but doubt has been cast on their general applicability by the finding that the evolved oxygen is sometimes in part reabsorbed owing to a secondary reaction with the blood in presence of ferricyanide. This can be overcome by taking adequate precautions.

(3) This error is avoided by combination of vacuum extraction and addition of ferricyanide. This principle has been developed by van Slyke to a high pitch of perfection, and his apparatus has been applied to numerous other estimations besides that of oxygen and carbon dioxide in blood. An acid ferricyanide solution and a known volume of the blood under examination, are run into an evacuated vessel ; the gases given off are extracted by shaking, and the residual solution is discharged by means of a two-way tap at the bottom of the extraction vessel. The volume of the extracted gases is determined either by transferring them to a burette, over mercury, and measuring the volume occupied at atmospheric pressure (constant pressure apparatus),

or by compressing them to a known volume over mercury, and measuring the pressure exerted, with a mercury manometer (constant volume apparatus). This is the more accurate method, since the scale on which the measurements are made is a great deal longer, and no errors can arise from faulty adjustment of the pressure of the gas in the burette. Carbon dioxide is absorbed from the extracted gases by adding a caustic soda solution, and oxygen by adding a solution of pyrogallol or sodium hyposulphite ; the contraction in volume (or reduction in pressure) is measured at each stage, and represents the volume of the respective gas present in the extracted mixture.

When plotted in terms of the oxygen content of the blood, the equilibrium curve rises, at large values of the oxygen pressure, to the value of the oxygen capacity : this depends on the hæmoglobin concentration of the blood, as measured, for example, by the hæmoglobinometer. When plotted in terms of the percentage saturation, the equilibrium curves of all samples of blood rise, of course, to the same maximum value. But such curves may be steeper or flatter, may be compressed or spread out along the axis of oxygen pressure, as the

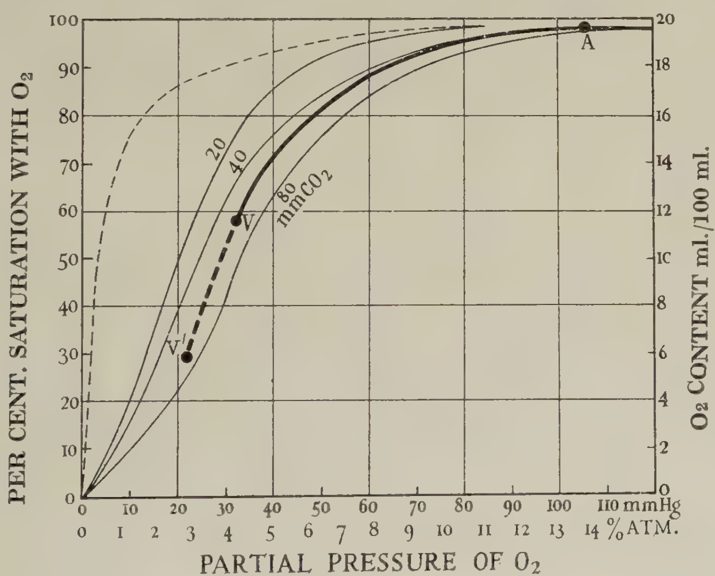


FIG. 3. 1. **Oxygen Equilibrium (Dissociation) Curves** of Blood at various Partial Pressures of Carbon Dioxide.

The scale of oxygen content on the right of the diagram may be used with reasonable accuracy for normal human blood. If the hæmoglobin concentration of the blood considered differs from the normal value, an appropriately different scale must be used.

AVV′ is the " physiological oxygen dissociation curve," A being the arterial point, V the venous point at rest, and V′ the venous point during exercise. Compare Fig. 3. 3.

The broken curve is a rectangular hyperbola ; note how little oxygen is given off until the partial pressure of oxygen is reduced to very low values. Such a curve is given by *myoglobin*, which will thus readily withdraw oxygen from hæmoglobin in the conditions of venous blood. (After Bock, Field and Adair, and R. Hill.)

temperature and the acidity of the blood are decreased or increased; and according to the kind of animal, or even particular individual, from which the blood was obtained. Increase of acidity, in particular, decreases the percentage saturation at a given oxygen pressure, *i.e.* decreases the affinity of the hæmoglobin for oxygen, and thus makes the curve flatter—an effect, as already remarked (p. 73), which results from the interaction between the hæm and the globin. Increased carbon dioxide pressure increases the acidity and hence has the same effect, as is shown by the family of curves at different carbon dioxide pressures given in Fig. 3. 1. (It is possible that carbon dioxide also exerts a specific effect due to its combining directly with hæmoglobin to a slight extent (see p. 91).) Lastly, the shape of the curve, as indicated by the size of the inflection at low pressures of oxygen, may depend on the electrolyte composition of the blood.

The dissociation curve of whole mammalian blood is always found to be S-shaped, as in Fig. 3. 1. This shape is of distinct physiological service, since it enables a large amount of oxygen to dissociate from the hæmoglobin without too severe a drop in the oxygen pressure with which it is in equilibrium. This is an important point as regards supply of oxygen to the tissues, for it is the partial pressure of oxygen in the blood, and not the amount of combined oxygen, that determines the rate of diffusion of oxygen from the blood to the tissues. A curve of the rectangular hyperbola type (*v.* broken curve in Fig. 3. 1) would, from this point of view, be obviously unserviceable.

The shape of the dissociation curve can be explained by applying the Law of Mass Action to the equilibrium between oxygen and hæmoglobin; but the reaction is somewhat complicated and is not accurately represented by the simple equation given on p. 73 above. Chemical analysis shows that the weight of hæmoglobin which contains 1 atomic weight of iron, *i.e.* 56 g., is 16,700 g. Direct determinations of the molecular weight of hæmoglobin from mammalian blood, by osmotic pressure determinations, or by the ultracentrifuge, give a value of 67,000. Thus each molecule of hæmoglobin contains 4 atoms of iron, and will combine with 4 molecules of oxygen. These can combine with the Hb_4 molecule one by one, forming the intermediate compounds Hb_4O_2, Hb_4O_4, Hb_4O_6, and finally Hb_4O_8. The S-shape of the dissociation curve is due to the fact that the affinity of each of the 4 hæm groups for oxygen depends on whether any or all of the other 3 have already combined with oxygen or not: in Hb_4O_6, in particular, the remaining unoccupied hæm must be presumed to have a much greater affinity for oxygen than it had when all 4 of the hæms were unoccupied. The 4 hæm groups must thus interact with each other, presumably through the globin part of the molecule: but since the magnitudes of all these interactions are not yet known, the Mass Action equation must contain 4 arbitrary constants. But if these are suitably chosen, the rather complicated equation fits the observed curves very accurately.

The effect of carbon dioxide (and acidity) on the affinity of hæmoglobin for oxygen is of considerable importance in enabling the blood to unload oxygen more readily into the tissues. It was first studied by Bohr, Hasselbalch and Krogh in 1904, and is often referred to as the **Bohr Effect.** Thus, for example, it can be seen from the curves that at an oxygen partial pressure of 32 mm. Hg the oxygen content of normal

human blood is 13 ml./100 ml. when the carbon dioxide pressure is 40 mm. Hg, and a little under 12 ml./100 ml. when the carbon dioxide pressure is increased to 50 mm. Hg. Now the partial pressure of carbon dioxide is about 40 mm. Hg in the arterial blood, and about 50 mm. Hg in the venous blood, so that roughly 1 ml. of extra oxygen is obtained by the tissues from each 100 ml. of blood, owing to this displacing action of acid, without lowering the partial pressure of oxygen within them.

Since carbon dioxide is blown off in the lungs, and taken up in the tissues, none of the family of curves shown in Fig. 3. 1 accurately represents the " physiological oxygen dissociation curve." This passes from a point A, corresponding to the conditions met with in the lungs, *i.e.*, 98 per cent. saturation with oxygen at a partial pressure of 108 mm. of oxygen and 40 mm. of carbon dioxide, to a point V, corresponding to the conditions met with in the tissues, *i.e.*, 58 per cent. saturation with oxygen at a partial pressure of 32 mm. Hg of oxygen and 50 mm. Hg of carbon dioxide.

It will be observed that the whole of the oxygen is never removed from the blood. This is due, first, to the fact that a small, but definite partial pressure of oxygen must exist in the tissue cells, and secondly, to the fact that oxygen must diffuse from the blood capillaries to the tissue cells, sometimes over quite a long distance. A considerable head of pressure is thus needed to drive the oxygen across at the requisite rate.

In exercise, the rate at which oxygen is needed by the muscles is greatly increased, and the amount of oxygen taken from each millilitre of blood is also increased, with the result that the partial pressure of oxygen in the venous blood is decreased (point V′ in Fig. 3. 1). There is thus a smaller head of pressure available for an increased rate of diffusion. This apparent contradiction is resolved by the observation that the number of capillaries carrying blood, per unit volume of muscle, is enormously increased when the muscle becomes active—possibly becoming 100 times greater. Consequently, not only is the distance decreased over which the oxygen must diffuse, but also the surface is increased over which oxygen can leave the blood.

The Fœtal Blood. The hæmoglobin in the blood of the fœtus has a greater affinity for oxygen than has that in the maternal blood. Its dissociation curve, even at relatively high partial pressures of carbon dioxide, would lie to the left, in Fig. 3. 1, even of that drawn for 20 mm. of carbon dioxide. When the maternal blood, therefore, is, say, 50 per cent. saturated with oxygen, the fœtal blood will be, perhaps, 80 per cent. saturated. Oxygen will consequently pass readily from the mother to the fœtus in the placenta. There is, of course, a corresponding disadvantage, in that the partial pressure of oxygen in the fœtal tissues must be small before the fœtal hæmoglobin will part with its oxygen.

Myoglobin. Most mammalian muscles contain a pigment which is closely related chemically to hæmoglobin, and is known as *myo-hæmoglobin*, or *myoglobin*. It differs from hæmoglobin in having an oxygen dissociation curve which is a rectangular hyperbola (broken curve in

Fig. 3. 1). Myoglobin will become nearly fully saturated with oxygen at partial pressures normally found in the tissues, and at which hæmoglobin has parted with most of its oxygen. It is known, also, that the tissue oxidation enzymes will function at oxygen pressures down to 5 mm. Hg or a little less, and at these pressures myoglobin will part with about 40 per cent. of its oxygen. Myoglobin, therefore, can act as an effective reservoir of oxygen in the muscles. This will be discussed further in Chapter 16.

Speed of the Reactions. Under physiological conditions the combination of oxygen with hæmoglobin, and the dissociation of oxyhæmoglobin only take about 1/100 second, and are thus too rapid to limit the rate of exchange of oxygen between the circulating blood in the capillaries (which each corpuscle takes about a second to traverse) and the lungs or tissues. Diffusion through the tissue cells and into the interior of the red blood cells seems to be the main factors which limit the rapidity of oxygen exchange in the animal.

Such very rapid chemical reactions cannot be timed by the ordinary methods—they require the special methods of Hartridge and Roughton for measurement of their rates. To determine the speed of combination, for instance, a solution of reduced hæmoglobin and a solution of oxygen in water are driven through separate leads into a small chamber where they mix in less than 1/1,000 second and travel thence into an observation tube. The percentage oxyhæmoglobin in the streaming fluid at various positions along the observation tube is measured spectroscopically, the fluid being kept in motion whilst the readings are taken. From the rate of flow of the liquid and the distance of the point of observation from the mixing chamber, the time taken by the reaction to reach the oxyhæmoglobin percentage recorded by the spectroscope is simply calculated, and hence the velocity of the reaction can be determined.

The Combination of Hæmoglobin with Carbon Monoxide. Carbon monoxide also combines reversibly with hæmoglobin, forming a compound usually known as *carboxyhæmoglobin*, and often written COHb. If a solution of hæmoglobin is equilibrated with a mixture of oxygen and carbon monoxide, the partial pressures, pO_2 and pCO being so large that no reduced hæmoglobin is present, the ratio of carboxyhæmoglobin to oxyhæmoglobin is defined by the equation :

$$\frac{\mathrm{COHb}}{\mathrm{O_2Hb}} = \mathrm{M}.\frac{\mathrm{pCO}}{\mathrm{pO_2}} \text{ where M is about 250.}$$

Thus, if the pressure of carbon monoxide is about 1/250 of that of oxygen, one-half of the hæmoglobin will be combined with carbon monoxide, and one-half with oxygen : hæmoglobin has about 250 times as great an affinity for carbon monoxide as it has for oxygen. This great affinity is due to the fact that carboxyhæmoglobin dissociates at least 1,000 times more slowly than does oxyhæmoglobin. In other respects, the reaction between carbon monoxide and hæmoglobin is very like that between oxygen and hæmoglobin : the volume of carbon monoxide combined at maximum saturation is the same as that of oxygen ; and

the dissociation curve (in the absence of oxygen) is also S-shaped and similarly affected by acidity and temperature.

The great affinity between carbon monoxide and hæmoglobin accounts for the danger attending the inhalation of small amounts of carbon monoxide or a mixture of gases containing it, *e.g.* coal gas. Blood can be completely saturated with carbon monoxide at a partial pressure of only 0·5 per cent. of an atmosphere at 37° C., whereas it requires at least 15 per cent. of oxygen before even approximate saturation is reached: if the whole of the available hæmoglobin becomes saturated with carbon monoxide, no oxygen can be carried, and the animal will be asphyxiated.

No marked symptoms are detectable until about 30 per cent. of all the hæmoglobin in the body is saturated with carbon monoxide. Vision, hearing and intelligence become impaired when the carbon monoxide saturation reaches 50 per cent., and death has been known to occur at 60 per cent. saturation; 80 per cent. saturation is almost invariably fatal. The average dissociation curve of carboxyhæmoglobin shows that with most people the first symptoms would be observed when the air they were breathing contained about 0·03 per cent. (1 part in 3,000) of carbon monoxide, while 0·4 per cent. (1 part in 250) would be fatal; different people, however, have somewhat different susceptibilities.

If an ordinary gas leak occurs in an ordinary room (a gas fire is turned on and not lighted, for example), it is very unlikely that there will ever be more than 2·5 per cent. of gas in the room, owing to leakage up the chimney and through the cracks of the window and diffusion through the walls and the ceiling to adjacent rooms (this would probably not apply to the case of a large gas cooker in a small kitchenette). Ordinary coal gas contains, as a rule, about 10 per cent. of carbon monoxide; water gas, however, is now usually added, and the gas as supplied may contain 20 per cent., or more, of carbon monoxide. There might thus be as much as 0·5 per cent. of carbon monoxide in the room, so that an ordinary leak in an ordinary room might have fatal consequences, and would almost certainly produce serious symptoms. Coal gas, it must be remembered, is lighter than air, and hence rises to the ceiling; a tall man may thus notice symptoms of anoxæmia before a short man.

Tobacco smoke contains appreciable quantities of carbon monoxide, and the deleterious effects of over-smoking have been attributed to a chronic anoxæmia produced by the continuous presence of carboxyhæmoglobin in the blood; it is, indeed, quite easy to detect the presence of this compound in the blood of a heavy smoker. The exhaust gas of motor cars also contains considerable amounts of carbon monoxide.

The treatment in all cases of carbon monoxide poisoning should be the administration of oxygen. The high pressure of oxygen not only facilitates the eventual dissociation of the carboxyhæmoglobin, but also enables an appreciable amount of oxygen to be carried in simple solution in the blood.

Carboxyhæmoglobin is cherry pink in colour rather than scarlet, like oxyhæmoglobin. Its absorption spectrum shows two bands in the visible spectrum, one in the yellow, the other in the green. These resemble the corresponding bands of oxyhæmoglobin, except that they are not quite so distinct and are nearer to the blue end of the spectrum. Unlike oxyhæmoglobin, carboxyhæmoglobin is very readily dissociated by exposure to strong light.

The estimation of carbon monoxide in blood is of importance both from the medico-legal and from the physiological points of view (*cf.* estimation of blood volume, Chapter 8); a variety of methods have been used:—

(1) *Optical.* (*a*) Depending on the difference in colour between carboxy-

hæmoglobin and oxyhæmoglobin (*cf.* Haldane's carmine titration method).

(*b*) Depending on the difference in position between the absorption bands of carboxyhæmoglobin and of oxyhæmoglobin (*cf.* Hartridge's reversion spectroscope).

(2) *Chemical.* Depending on the colour of the precipitate formed by COHb with tannic acid.

(3) *Gasometric.* Depending on an adaptation of van Slyke's methods for oxygen and carbon dioxide estimation in blood.

Where sufficient blood is available, the gasometric method is the most accurate and reliable.

The Regulation of the Hydrogen Ion Concentration

When oxygen is shaken with water, it merely goes into physical solution, and does not form any chemical compound, unless there be added to the water some substance like hæmoglobin, which is specially capable of combining with the oxygen. Carbon dioxide, on the other hand, not only dissolves readily in water (or any watery fluid), but also reacts chemically with the water to form carbonic acid, H_2CO_3; this being a weak acid, splits up to some extent into hydrogen ions and bicarbonate ions. Bicarbonate ions are also capable of splitting up further into hydrogen ions and carbonate ions, but this second ionisation only occurs appreciably at more alkaline reactions than those found in the body fluids.

The first ionisation, viz., the formation of H^+ and HCO_3^- ions, is, however, very marked under the conditions in the body and hence the blood would go markedly acid when it takes up carbon dioxide in the tissues, were there not present some chemical mechanism to counteract such changes, namely, the buffer mechanism explained later. The presence of the latter is hardly to be wondered at, for many processes in living organisms, such as the velocity of enzyme actions, and the stability of proteins, are greatly affected by changes in acidity and to *different extents*; hence any marked change in acidity would tend to throw the whole organism out of gear.

Carbon dioxide transport and maintenance of neutrality are both important, and both closely interconnected: for a clear understanding of them some knowledge of the general physical chemistry of buffer systems and of ionic equilibria in solution is needed. The reader is recommended to consult a suitable text-book of biochemistry or physical chemistry.

Buffer Action. If we titrate a weak acid, such as carbonic acid, with a strong base, such as sodium hydroxide, and plot the pH of the solution against the number of gram-molecules of base added, we get a curve such as that shown in Fig. 3. 2. Clearly, the steeper is the titration curve, the more base (or, of course, acid) must be added in order to produce a given change in pH, and the more strongly is the solution said to be *buffered.* The salt of the weak acid is very nearly completely ionised, while the free acid is almost completely un-ionised; we can thus regard the variation with pH of the amount of base added, as equivalent to the variation with pH of the degree of ionisation of the

acid. The maximum slope, and the strongest buffer action, occurs when the weak acid is exactly half titrated, *i.e.*, when the hydrogen ion concentration is equal to the apparent dissociation constant of the weak acid. In practice, the region of hydrogen ion concentration over which the buffering power is reasonably large is taken to be ten times greater to ten times less than the value of the dissociation constant. Alterna-

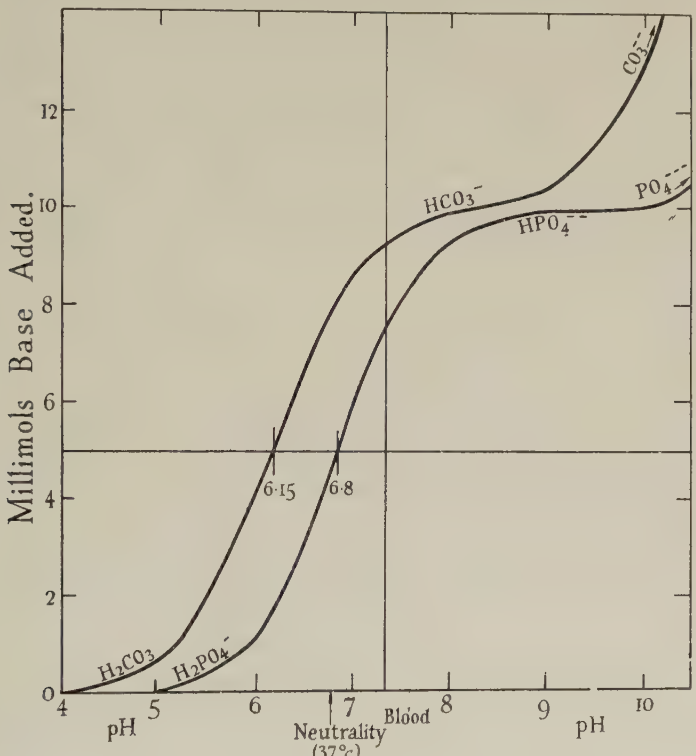


FIG. 3. 2. Titration Curves of Weak Acids.

H_2CO_3 and NaH_2PO_4 in 10 mM. solution are titrated with NaOH. The value of pK′ is given by the pH at which the acid is half titrated. This value varies with the nature and concentration of the other ions in the solution ; the curves are drawn approximately for the conditions in plasma. The vertical line is drawn at the pH of blood. (After L. J. Henderson.)

tively, we can say that the pH must be within one unit on either side of the value of pK′. The buffering power of a given solution is also, of course, directly proportional to the concentration of those buffer substances whose pK′ is within one unit of the pH of the solution. Exactly similar arguments can be used in connection with weak bases, when titrated with strong acids ; they also act as buffers, but are not met with in any appreciable concentration in solutions of physiological interest.

The reader may perhaps be reminded that the pH is a convenient measure of the hydrogen ion concentration, which is usually extremely small. The pH is defined as the negative logarithm of the hydrogen ion concentration; so that a hydrogen ion concentration of, say, 5×10^{-8}, or $10^{-7 \cdot 3}$ g. ions per litre corresponds to a pH of 7·3.

The shape of the titration curves may be deduced from the Law of Mass Action. If we have a weak acid HA which dissociates into H^+ and A^-, the following reaction will be in equilibrium.

$$HA \rightleftharpoons H^+ + A^-$$

whence
$$[H^+] = K_a . \frac{[HA]}{[A^-]}$$

or
$$pH = pK_a + \log \frac{[A^-]}{[HA]}$$

square brackets denoting concentrations. These equations indicate that so long as the H^+ ion concentration is of the same order of magnitude as the dissociation constant K_a, the greater is the H^+ concentration, the less is the acid ionised.

The buffer substances present in blood are (*a*) bicarbonate, (*b*) hæmoglobin, (*c*) the plasma proteins, and (*d*) phosphates. From Fig. 3. 2 we see that the bicarbonate system is not, by itself, a very effective buffer at the normal pH of blood—its pK value is 1·2 units from the pH of blood. It is present, however, in relatively high concentration, so that its effect is quite appreciable. The $Na_2HPO_4 - NaH_2PO_4$ system, on the contrary, is effective as far as buffer power is concerned, but the concentration of phosphates is so low that they play little part in buffering the blood. In the urine, phosphates provide the greater part of the buffering in normal circumstances. The most important buffering substance in the blood is undoubtedly hæmoglobin; this, in the corpuscles, exists as a mixture of the potassium salt KHb, which may be regarded as ionised into K^+ and Hb^-, and the free acid HHb, which is un-ionised. Hæmoglobin is a complex polybasic acid, and its titration curve consists of a large number of the S-shaped curves of Fig. 3. 2 on top of one another, and overlapping. The general conception of buffer action is nevertheless still applicable. The plasma proteins behave in a similar way, but are present in much smaller concentration. If, then, we add a small quantity of an acid to some blood, the hydrogen ion concentration will be increased. This, however, will lead to a reduction in the ionisation of both hæmoglobin and carbonic acid, and many of the extra hydrogen ions will be absorbed and tucked away in the undissociated acids. Conversely, if an alkali be added, the extra hydroxyl ions will combine with hydrogen ions to form water, and the effect will be exactly the opposite of that which occurs when acid is added. Hæmoglobin and carbonic acid will dissociate more completely, and provide extra hydrogen ions to replace those which were removed by the alkali. All this is, of course, but a verbal description of the fact illustrated in the titration curves, that in a buffered system the pH changes only slowly with the addition of acid or alkali.

The efficiency of the buffering process in the blood is indicated by

the following calculation. If we take blood at a pH 7·4 and add acid until it has a pH 7·2, the increase in free hydrogen ion concentration is $10^{-7\cdot2} - 10^{-7\cdot4}$ or $2\cdot3 \times 10^{-8}$ g.-ions per litre. During the course of this change in pH, however, the hæmoglobin absorbs, by buffer action, no less than $660{,}000 \times 10^{-8}$ g.-ions per litre, while the carbonic acid absorbs $70{,}000 \times 10^{-8}$ g.-ions per litre. The total amount of hydrogen ions which must be added to produce this change in pH is thus $730{,}000 \times 10^{-8}$ per litre of blood, of which only $2\cdot3 \times 10^{-8}$ remain free. If we remove hydrogen ions from normal blood (or add hydroxyl ions) until the pH is 7·6, the decrease in free hydrogen ion concentration is $1\cdot5 \times 10^{-8}$ g.-ions per litre, while hæmoglobin releases again $660{,}000 \times 10^{-8}$ g.-ions per litre and carbonic acid releases $44{,}000 \times 10^{-8}$ g.-ions per litre. These figures indicate the relative importance of carbonic acid and hæmoglobin in buffering the blood. Roughly, 90 per cent. of the hydrogen ions added or removed are absorbed or released by hæmoglobin, and 10 per cent. by the carbonic acid-bicarbonate system.

The above calculations are based on the following equations : (1) The buffering power of hæmoglobin is represented by the empirical equation

$$[HbO_2^-] = 0\cdot22\,[HbO_2]\,(pH - 6\cdot60)$$

$[HbO_2^-]$ is expressed in milli-equivalents per litre and $[HbO_2]$ in grams per litre.

(2) The buffering power of the carbonic acid-bicarbonate system is represented by the well-known Henderson-Hasselbalch equation

$$pH = 6\cdot12 + \log[HCO_3^-] - \log[H_2CO_3]$$

The values taken are :

$[HbO_2]$ 150 g./litre.
Total $CO_2 = [H_2CO_3] + [HCO_3^-] = 24\cdot8$ millimoles/litre.
Total base $= [HCO_3^-] + [HbO_2^-]$ at pH 7·4 $= 50$ millimoles/litre.

In this example, we have imagined that no carbon dioxide is lost or gained by the blood when the pH is changed. This approximates to the conditions in the capillaries when some acid product of metabolism is formed in the tissues and diffuses into the blood stream. If we consider the body as a whole, however, the conditions are different, since carbon dioxide can be blown off in the lungs and acid or base can be excreted by the kidneys. As we shall see in later chapters, the respiratory centre and the kidneys adjust the ventilation rate and pH of the urine, respectively, in such a way as to keep the pH of the blood constant. The buffering is thus perfect. Confining ourselves for the moment to the actions of the respiratory centre, we see that the buffering results, in the end, entirely from the increased loss, or retention, of carbon dioxide in the lungs. The extra hydrogen ions are entirely absorbed, or released, by the bicarbonate system, and hæmoglobin in the long run plays no part. There are thus three lines of defence against changes in hydrogen ion concentration : (1) direct buffering in the blood, mainly by hæmoglobin ; (2) indirect buffering by the action of the respiratory centre in controlling the free carbon dioxide concentration in the blood ; and (3) indirect buffering by the kidneys, which more slowly excrete the excess acid or base, as the case may be.

The Carriage of Carbon Dioxide

Blood loses oxygen and gains carbon dioxide in its passage through the tissues and undergoes the reverse changes during its passage through the lungs. In the tissues and lungs the changes are roughly equal and opposite, so that it will only be necessary to describe one of them. For convenience, we shall consider the uptake of carbon dioxide by the blood in the tissues. The partial pressure of carbon dioxide in arterial blood is usually about 40 mm. Hg, and the total carbon dioxide content (as estimated by vacuum extraction with acid) is about 50 ml. (at N.T.P.) per 100 ml. The partial pressure of carbon dioxide in the tissues is higher than that in the arterial blood, hence carbon dioxide diffuses from the tissues into the blood capillaries. The amount of carbon dioxide taken up by the blood is some twenty times greater than the amount taken up by water under like conditions of carbon dioxide pressure, hence only about 5 per cent. of the carbon dioxide uptake by the blood is by simple solution, the remaining 95 per cent. being through chemical combination.

The problem then, as in the case of oxygen, is to discover the means whereby each unit volume of blood takes up in the tissues, and gives out in the lungs, so much more carbon dioxide than would water in similar circumstances. We now know that there are two chief means by which this is done ; one resulting from the combination of carbon dioxide with water, to form carbonic acid, which is buffered just as is any other acid ; and the other resulting from a direct combination of carbon dioxide with hæmoglobin.

In forming our ideas as to the transport of carbon dioxide, we must distinguish between the results of experiments in which blood is shaken outside the body with various gas mixtures, and the behaviour of the blood in the circulation. The former usually takes about fifteen minutes or longer ; in the latter, the blood is often only in contact with the tissues (where it takes up carbon dioxide) and with the lungs (where it gives up carbon dioxide) for times of the order of one second. Attention must therefore be paid to the speed of the various processes, and this has revealed the presence of important factors which would otherwise have been missed.

Carbon Dioxide as an Acid. We have already discussed, in the previous section, how carbon dioxide combines with the water in which it is dissolved, to form carbonic acid ; and how the buffer substances in the blood (chiefly hæmoglobin) prevent the acidity from changing to any serious extent. The reactions involved may be written :

$$CO_2 + H_2O \rightleftharpoons H_2CO_3$$
$$H_2CO_3 \rightleftharpoons H^+ + HCO_3^-$$
$$H^+ + Hb^- \rightleftharpoons HHb$$

These reactions can be combined into the one equation :

$$CO_2 + H_2O + Hb^- \rightleftharpoons HCO_3^- + HHb$$

The net result of the series of reactions therefore is that almost all the extra carbon dioxide is carried in the blood in the form of HCO_3^- ions, the negative charges being supplied by the Hb^- ions. By this means the amount of carbon dioxide that can be carried in a given

volume of blood is increased enormously over that which would be carried at the same partial pressure, in simple solution.

This does not, strictly, express the whole story. There must be some increase in hydrogen ion concentration, otherwise there would be no reason why the ionisation of the hæmoglobin should be depressed. This rise in hydrogen ion concentration has the effect of preventing some of the added carbon dioxide from being converted into bicarbonate ions. A little, therefore, remains as carbonic acid (the partial pressure of the carbon dioxide in venous blood is greater than that of arterial blood); the majority forms hydrogen and bicarbonate ions, and most of the hydrogen ions are absorbed by the hæmoglobin.

The reaction of carbon dioxide with the plasma proteins may be expressed in a similar way, *viz.*:

$$H_2CO_3 + Pr^- \rightleftharpoons HCO_3^- + HPr$$

and with phosphates by the equation:

$$H_2CO_3 + HPO_4^{2-} \rightleftharpoons HCO_3^- + H_2PO_4^-$$

The plasma proteins contribute only about 7 per cent., and the phosphates less than 3 per cent. of the total buffering and carbon dioxide carrying power of the blood. They are, therefore, of minor importance as compared with hæmoglobin.

Oxyhæmoglobin and Reduced Hæmoglobin as Acids. This superior efficiency of hæmoglobin over the plasma proteins is in part due to its higher concentration and its greater ionisation in the physiological pH range; but it is, in the main, due to a much more important factor which has not yet been considered. In the living animal, when blood takes up carbon dioxide it also loses oxygen from combination with hæmoglobin, so that the latter is left in the reduced form. Now it has been definitely shown that reduced hæmoglobin is a *weaker acid* than oxyhæmoglobin, *i.e.*, at a given pH, reduced hæmoglobin is less ionised than oxyhæmoglobin. This means that the reaction

$$H_2CO_3 + Hb^- \rightleftharpoons HCO_3^- + HHb$$

must proceed further to the right at a given partial pressure of carbon dioxide, than does the reaction

$$H_2CO_3 + HbO_2^- \rightleftharpoons HCO_3^- + HHbO_2$$

i.e., the reduced hæmoglobin ions do not hold on to their negative charges so firmly as do the oxyhæmoglobin ions, and so transfer them more readily to the bicarbonate ions. Reduced blood, therefore, takes up carbon dioxide in the tissues more readily than does oxygenated blood. Conversely, in the lungs when oxyhæmoglobin is re-formed from reduced hæmoglobin the carbon dioxide will be liberated from the blood more readily.

These relationships are well shown by a comparison of the CO_2 dissociation curves of oxygenated and reduced blood. In curves of this kind the total concentration of carbon dioxide (chemically combined and dissolved) is plotted against the partial pressure of carbon dioxide in the gas phase when equilibrium has been reached with the blood. In

Fig. 3. 3 the point A represents the average condition of the arterial blood, V the average condition of the mixed venous blood at rest, V′ the condition of the venous blood in exercise. V and V′ are both situated between the oxygenated and reduced blood curves, since in the circulation the venous blood is only partially reduced. It will be seen that the line AV is about twice as steep as the carbon dioxide dissociation curve of oxygenated or reduced blood : the change from the fully oxygenated to the partially reduced state enables the blood to take up about twice as much carbon dioxide for a given increase of CO_2 pressure as it would

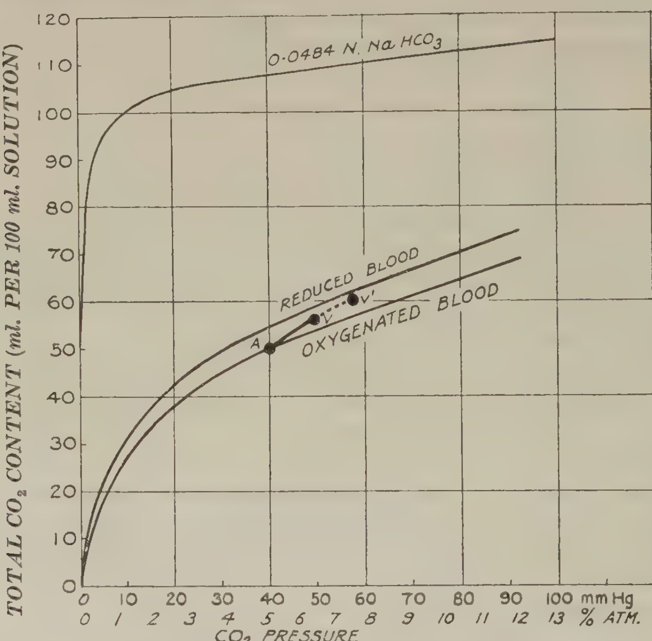


FIG. 3. 3. **Carbon Dioxide Equilibrium (Dissociation) Curves** of Oxygenated and Reduced Blood, and of a Solution of Sodium Bicarbonate of the same concentration of Total Available Base as the Blood.

AVV′ is the " physiological carbon dioxide dissociation curve," A being the arterial point, V the venous point at rest, and V′ the venous point during exercise. Compare Fig. 3. 1. (After Parsons.)

if there were no change in the strength of the hæmoglobin as an acid. This phenomenon, first studied by Christiansen, Douglas and Haldane in 1913, is sometimes known as the **Haldane Effect.**

The whole difference in CO_2 carrying power of oxygenated and reduced blood must not be attributed to the difference in acid strengths of oxy- and reduced hæmoglobin. The formation of carbamino compounds (see p. 91) is also partly responsible.

The explanation of how reduced hæmoglobin comes to be a weaker acid than oxyhæmoglobin is probably that one of the $-NH_2$ groups in the molecule, which takes up H^+ ions, is close to the hæmatin nucleus, and that when O_2 combines with the latter, it reduces the ease of binding of the H^+ ion

at this neighbouring $-NH_2$ group. Oxyhæmoglobin, being thus less able to bind H ions, behaves as a stronger acid.

The fact that loss of oxygen from the blood, as it passes through the tissues, renders it *ipso facto* more competent to take up carbon dioxide, just where such extra power is needed, and conversely, the fact that the uptake of carbon dioxide *ipso facto* drives off oxygen into the tissues, are beautiful examples of the way in which divers chemical phenomena are co-ordinated into a harmonious physiological process.

The "Chloride Shift." When carbon dioxide enters the blood from the tissues, some stays in the plasma, and the rest passes into the corpuscles. In both media the carbon dioxide combines with water to form carbonic acid, which dissociates into hydrogen ions and bicarbonate ions. Owing to the far greater buffering power of the corpuscle contents, this dissociation will proceed much more extensively in the corpuscle than in the plasma. Consequently, there will be a concentration gradient propelling carbon dioxide (undissociated) from plasma to corpuscle, and another gradient propelling bicarbonate ions from corpuscles to plasma. But, as discussed more fully in Chapter 8, the membranes of the corpuscles, although permeable to anions such as chloride or bicarbonate, behave as if they were impermeable to cations such as sodium or potassium. Bicarbonate ions will begin to diffuse out under the influence of the concentration gradient, but since they cannot be accompanied by ions of the opposite charge, the corpuscle will immediately develop a net positive charge. The positively charged corpuscle will then draw in negative ions of all kinds from the plasma. Chloride ions, being the most readily available, will predominate, and the process will continue until finally an equilibrium is reached at which

$$\frac{[HCO_3^-] \text{ in corpuscle}}{[HCO_3^-] \text{ in plasma}} = \frac{[Cl^-] \text{ in corpuscle}}{[Cl^-] \text{ in plasma}}$$

this being a case in which the physico-chemical principle known as the Donnan membrane equilibrium applies. The net result is that when the carbon dioxide content of the blood is increased, there is a migration of bicarbonate ions out of the corpuscles into the plasma, and a migration in exchange of chloride ions from the plasma into the corpuscles. When the carbon dioxide content is reduced, chloride ions come out of the corpuscles into the plasma, and bicarbonate ions enter instead. The carbon dioxide carrying power of the plasma is thus brought up to the level of, or even beyond that, of the corpuscles. Since the plasma hydrogen ion concentration is determined by the ratio of the carbon dioxide pressure to the concentration of bicarbonate ions, the transfer of bicarbonate ions from the corpuscles will assist in preventing the hydrogen ion concentration of the plasma from rising. The superior buffering power of the corpuscles is thus shared out with the inferior buffering power of the plasma—a process often spoken of as "secondary buffering" of the plasma: it is accompanied by the chloride shift only because the red cell membrane is sensibly impermeable to sodium and potassium ions.

If the plasma is replaced by isotonic NaCl solution, the chloride shift again occurs when the blood corpuscle suspension in NaCl is shaken with carbon dioxide, and for the same reasons as above. If, however, isotonic sugar solution is used, there is now no ion in the outside fluid to exchange with HCO_3^- from the corpuscle, so that when the corpuscle suspension is shaken with carbon dioxide, the outside sugar solution cannot be secondarily buffered, and, since it contains no intrinsic buffer, it therefore goes very acid.

The chloride shift is mainly responsible for the difference between the carbon dioxide dissociation curve of " separated plasma " and that of " true plasma." The curve for " *separated plasma* " is obtained by centrifuging the blood, removing the supernatant fluid, shaking the latter with various pressures of carbon dioxide and then estimating the total CO_2 content at each CO_2 pressure. Over the physiological range of carbon dioxide pressure the curve is almost as flat as the CO_2 dissociation curve of a solution of sodium bicarbonate (Fig. 3. 3) : such a curve would be of little physiological service, since the extra amount of carbon dioxide which would be taken up when its pressure is raised from 40 to 50 mm. Hg would be so small. The dissociation curve of " *true plasma* " is obtained by equilibrating the *whole* blood with various pressures of carbon dioxide, and then transferring the blood to a centrifuge cup (the escape of carbon dioxide into the air is prevented by a layer of liquid paraffin) ; the supernatant fluid, separated by centrifuging, is estimated for total CO_2, and the latter plotted against CO_2 pressure. Each sample of plasma thus is in equilibrium, not only with each different pressure of CO_2, but also with the corpuscles—whereas all the samples of separated plasma were in equilibrium with the corpuscles at one particular CO_2 pressure, namely, that which happened to obtain in the blood at the time when the plasma was centrifuged off. The true plasma dissociation curve is clearly quite different from the separated plasma curve : it is, indeed, of the same type as that of whole blood, and like the latter is of a serviceable shape as regards carbon dioxide transport.

As a result of all these mechanisms, the hydrogen ion concentration of the plasma is kept remarkably constant in the neighbourhood of pH 7·4 when the body is at rest. The difference in pH between arterial and venous blood at rest is only about 0·03, but in work it is rather greater, viz., about 0·08.

The Speed of the Processes involved in Carbon Dioxide Transport. Most of the individual chemical reactions we have described as occurring in the uptake or output of carbon dioxide are of an ionic type, *e.g.*,

$$H_2CO_3 \rightleftharpoons H^+ + HCO_3^-,$$
$$H_2PO_4^- \rightleftharpoons H^+ + HPO_4^{--},$$
$$\text{HPr (protein molecule)} \rightleftharpoons H^+ + Pr^-.$$

Simple ionic reactions of this type have been generally supposed to be very rapid—experiments have, indeed, shown that all these reactions reach to within 1 per cent. of equilibrium within 1/1,000 second (from whichever side of it they start) ; they must therefore be too fast to limit the rate of output or uptake of carbon dioxide by the blood. The final

chemical reaction, which, on the bicarbonate hypothesis, comes just before the evolution of carbon dioxide into the expired air is, however, a non-ionic one, viz., the formation of CO_2 from H_2CO_3.

The value of the velocity constant of this reaction has been measured by several physical chemists : it can thence be calculated that at body *p*H and temperature, and with the usual concentration of bicarbonate ions in the blood, carbon dioxide could only be evolved in the lungs at about 1/200th the rate at which it actually escapes into the expired air. Either, then, there must be something in the blood which speeds up the reversible reaction between carbon dioxide and water ; or else there must be some other reaction, besides the bicarbonate one, which takes part in the transport of carbon dioxide : there are, in fact, both.

Carbonic Anhydrase. It can be shown, by quite simple experiments, that the red blood corpuscles contain a substance which catalyses strongly the reaction $H_2CO_3 \rightleftharpoons CO_2 + H_2O$. This substance has been isolated, was found to have all the properties of an enzyme, and was thus given the name " carbonic anhydrase." It is not present in the plasma, but the amount in the corpuscles, if it is as efficient there as in solution, is enough to accelerate the formation of CO_2 from H_2CO_3 about 600 times under body conditions. Presumably during the short time the circulating blood is in the capillaries, the change from CO_2 to bicarbonate and *vice versa* must occur chiefly in the red corpuscles.

Carbhæmoglobin. Blood can still react *rapidly* with a small amount of carbon dioxide even when carbonic anhydrase is absent or incapacitated by addition of some enzyme poison, *e.g.*, KCN ; this residual rapid reaction cannot be bicarbonate formation or the reverse (since both these processes, in absence of carbonic anhydrase, only proceed slowly). There is chemical evidence that it is due to a direct reaction of CO_2 with the $—NH_2$ groups of hæmoglobin to form compounds of a carbamino type, *e.g.*,

$$Hb(NH_2) + CO_2 \rightarrow Hb(NHCOO^-) + H^+$$

This type of reaction is well known in the case of simpler $—NH_2$ containing compounds, such as ammonia and glycine, and is a very rapid one even in the absence of special catalysts.

Carbamino bound carbon dioxide plays an appreciable *rôle* in carbon dioxide transport, even though the absolute amounts of such compounds under physiological conditions form only a small fraction of the total amount of carbon dioxide present in the blood (in the neighbourhood of 5 per cent.). Reduced hæmoglobin takes up carbon dioxide in the carbamino form more readily than does oxyhæmoglobin ; and of the total quantity of carbon dioxide carried from the tissues to the lungs—*i.e.*, of the difference between the carbon dioxide contents of venous and arterial blood—it is estimated that about one-quarter is in the carbamino form.

Other evidence, of a more physico-chemical type, suggests that some carbon dioxide combines directly with hæmoglobin, not only in the carbamino form, but possibly in some other form.

Summary. The uptake of carbon dioxide by the blood in the tissues is believed to occur as follows :—

(1) CO_2 diffuses from the tissues into the blood plasma.

(2) Some of the CO_2 hydrates slowly in the plasma to form H_2CO_3 ; the latter then yields its H^+ ions to the plasma proteins and phosphates and forms bicarbonate ions.

(3) Most of the CO_2, however, passes into the red corpuscles :

(*a*) Some combines directly with hæmoglobin to form compounds of a carbamino type—this combination is increased as the hæmoglobin loses oxygen in the blood capillary.

(*b*) By far the greater part changes over rapidly into H_2CO_3 under the influence of the enzyme carbonic anhydrase ; the H_2CO_3

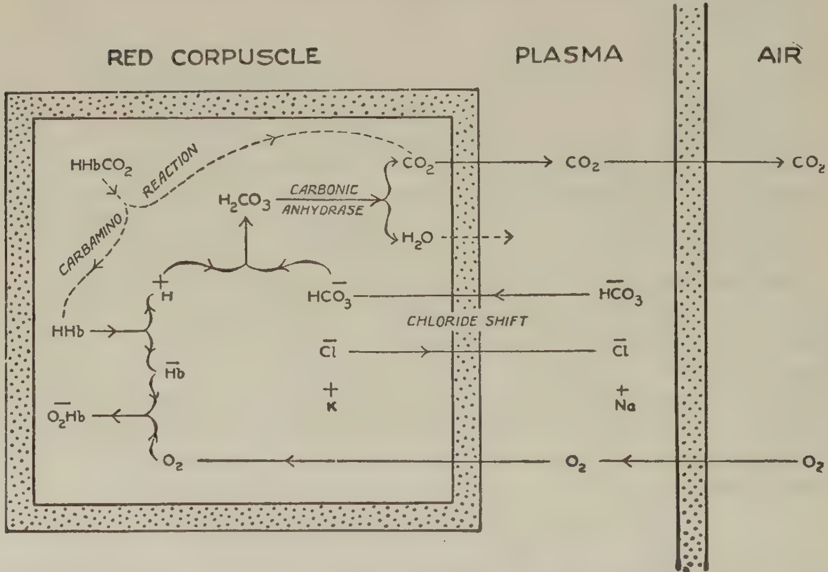


FIG. 3. 4. Scheme showing the Most Important Changes involved in the Liberation of Carbon Dioxide from Blood into Air. (Roughton.)

then yields its H^+ ions to the hæmoglobin, forming bicarbonate ions. The latter process is much increased as the hæmoglobin changes from the oxy- to the reduced form, since reduced hæmoglobin, being a weaker acid than oxyhæmoglobin, absorbs H^+ ions more readily and thereby allows H_2CO_3 to become more completely ionised.

(4) Bicarbonate ions begin to diffuse out from the corpuscles into the plasma ; this sets up an electric field which draws Cl^- ions into the corpuscles from the plasma in place of the HCO_3^- ions, which have diffused out (the chloride shift). The twin process goes on until an equilibrium is reached. The CO_2 carrying power and buffer efficiency of the plasma are thereby brought up to the level of the red corpuscles.

In the lung carbon dioxide is formed and evolved by a reversal of all these processes, as indicated in Fig. 3, 4.

The Effect of Exercise

The changes that take place in the blood as a result of exercise, or of increased activity of any kind, are quantitative rather than qualitative, and their magnitude depends largely upon the magnitude of the simultaneous changes in the respiratory and circulatory systems.

The first result of an increased activity in any group of cells is that more oxygen is used and more carbon dioxide evolved ; the partial pressure of oxygen around them falls, and that of carbon dioxide rises. This effect is passed on to the blood in the capillaries, so that the quantities defining the venous points on our dissociation curves are altered, taking up, for example, values somewhat as shown in Figs. 3. 1 and 3. 3 by the points marked V′, *i.e.*, the oxygen pressure falls to 22 mm. Hg, the carbon dioxide pressure rises to 58 mm. Hg, the per cent. saturation with oxygen falls to 30, and the total carbon dioxide concentration rises to 60 ml. per 100 ml. The extra carbon dioxide is thus carried away, and the extra oxygen provided, simply by allowing the partial pressure in the blood of the one to rise and of the other to fall. We have seen that this does not necessarily involve similar changes in the tissue cells, owing to the opening up of extra capillaries and the reduction in the distance that has to be traversed by the gases between the blood and the tissue cells. The increased blood-flow required is provided for by the vasomotor reflexes and local chemical mechanisms as described in Chapter 2.

A limit must come to this, however, and if the activity continues to become greater, the oxygen pressure in the tissue cells begins to fall, and the oxygen supply to be deficient. The cells go into "oxygen debt" (see Chapter 16), and, if this is severe enough, begin to liberate lactic acid. This addition of extra hydrogen ions drives the equilibrium reaction,

$$H_2CO_3 \rightleftharpoons H^+ + HCO_3^-$$

from right to left. The carbonic acid so formed is converted to carbon dioxide and blown off in the lungs (the respiratory centre adjusts the rate of ventilation so as to ensure this). The buffering power of the tissues is reduced, nevertheless, since the buffer salt, bicarbonate, is replaced by the fully ionised lactate, which has no buffer action ; the hydrogen ion concentration of the tissues rises. If the exercise continues, there may be so much lactic acid produced that the hydrogen ion concentration of the arterial blood is affected to a significant extent. We shall see in the next chapter that the carbon dioxide pressure in the lungs may rise also, and these two effects may so alter the oxygen dissociation curve that the per cent. saturation of the arterial blood falls. The oxygen and carbon dioxide transport system has now got into a vicious circle, and the exercise must stop. It is, indeed, sometimes terminated by unconsciousness.

It should be remarked that the whole of the body is affected by the changes that result in the blood from one group of muscles going into oxygen debt ; the lactic acid, indeed, diffuses into all the wet tissues of the body, and may be oxidised in them—and particularly in the liver and heart—as well as in the muscles in which it was formed. The blood buffers play somewhat the part of a bank which allows one of its clients to overdraw his account at the expense of the rest, but if the overdraft is allowed to increase until it is comparable in size with the whole of the negotiable securities of the bank—and this in spite of the activities of the other clients to redeem it—the bank's failure is inevitable.

CHAPTER 4

RESPIRATION

CONTINUED vital activity of tissues demands the appropriate interchange of the gases, oxygen and carbon dioxide, between the tissues and the atmosphere. In its more usual sense, respiration means the operation of the special apparatus concerned with the absorption of oxygen by, and the removal of carbon dioxide from, the body as a whole ; this is termed external respiration. Internal, or tissue respiration, on the other hand, is the local process of utilisation of oxygen and production of carbon dioxide by the tissue cells, and will be considered in a later chapter on Metabolism. In the present chapter we are mainly concerned with the processes of external respiration.

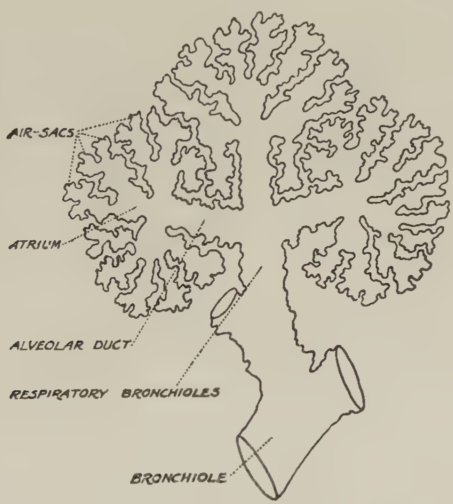


FIG. 4. 1. Sectional diagram of a Structural Unit of the Lungs. (After Miller.)

The Structure and Movements of the Respiratory Mechanism

Air enters and leaves the lungs by traversing, first the nasal passages (and also the buccal cavities when the mouth is open), where the vascular mucous membrane warms and moistens the incoming air ; next the pharynx, the larynx, and lastly the trachea into the bronchi. The final branching network of the bronchial tree opens into the terminal air sacs with their alveolar saccules, or alveoli (see Fig. 4. 1). It is in the alveoli that the respiratory gases, oxygen and carbon dioxide, are able to make close contact with the blood in the capillaries.

The tubes of the trachea and bronchi are kept permanently open by a series of cartilaginous rings embedded in their walls. These are partly made up of a dense fibro-elastic membrane. The whole of the bronchial tree is richly supplied with elastic fibres, mainly disposed in a longitudinal direction. These fibres, along with the elastic tissues in the lungs, account for the recoil of the lungs and the bronchial tree which takes place during expiration (see Fig. 4. 2).

The respiratory passages are lined with ciliated and with mucus-secreting cells. The cilia produce a constant wave-like motion in the direction of the nasal and buccal cavities and are very efficient in expelling any foreign material that may come to rest on their surfaces.

The pharynx is a common pathway for air and food. In order to prevent the food going down the " wrong way," the aperture of the

Fig. 4. 2. X-ray photographs (retouched) of the **Bronchial Tree** of a young woman : (A) in full expiration, (B) in full inspiration. (Macklin.)

larynx is guarded by the epiglottis. During the process of swallowing, the arytenoid cartilages are closely approximated and pulled forwards towards the epiglottis, and at the same time breathing is inhibited (deglutition apnœa).

The lungs are so constructed that an almost instantaneous exchange of gases can take place between the air within them and the blood passing through them. The pulmonary capillaries, although individually only about 8μ in length, may altogether expose a surface area to the lung gases of approximately 70 sq. metres ; the blood in these capillaries is exposed for no more than half a second. Interposed between the blood and the air in the alveoli are two delicate membranes each of one cell thickness, namely, the epithelium forming the alveolar wall and the endothelium of the capillary. Across such a membrane the respiratory gases can readily diffuse.

The Bronchial Muscles

The walls of the bronchi and bronchioles contain strips of plain muscle which run spirally round them forming a " geodesic " network. Contraction of this muscle, therefore, both narrows the bore and decreases the length of these tubes, and consequently reduces the volume of air contained in the lungs at any given value of intrapleural pressure, *i.e.* increases the elastic recoil. The narrowing of the bore also increases the resistance to the passage of air in and out of the lungs and causes laboured breathing when the contraction is excessive.

Blood Supply. The whole of the bronchial tree, as far as the respiratory bronchioles, is supplied by the bronchial arteries which are branches of the aorta. The blood is collected by bronchial veins and eventually drains into the right atrium; part of the blood is returned to the left atrium via the pulmonary veins.

Nerve Supply. The *efferent* nerve supply is derived entirely from the autonomic nervous system. The vagus nerves are constrictor to the bronchioles whereas stimulation of the sympathetic nerves dilates them. The bronchioles also respond to the action of autonomic drugs; parasympathomimetic ones, such as acetylcholine or pilocarpine, constrict, and sympathomimetic drugs, adrenaline and isoprenaline, dilate them (Chapter 15). There is also some evidence that the calibre of the bronchi and bronchioles increases during inspiration and diminishes again on expiration, thereby assisting the diaphragm and the intercostal muscles in renewing the air in the lungs. The mechanism for this is largely passive.

The *afferent* nerve supply to the laryngeal mucosa is the superior laryngeal nerve. Stimulation of its nerve endings, by the presence of a foreign body or as a result of disease, reflexly causes coughing. From the bronchial tree and the lung, afferent fibres run up in the vagus nerves ; their chief function is to indicate to the respiratory centre the degree of expansion of the lungs. This will be referred to later.

Asthma. Under certain conditions the muscles in the bronchioles are stimulated by irritation in various parts of the body, especially the nose and the air passages themselves, and undergo spasmodic contraction. This results in the condition known as asthma, in which great difficulty is experienced in breathing, particularly in expiration, since this is normally a passive movement. Bronchial asthma is associated also with hypersensitivity to certain proteins, notably those in the pollens of some grasses. This is analogous to the sensitivity of an anaphylactic type responsible for urticaria and hay fever (see Chapter 21). Anaphylaxis in general appears to be associated with the production of a histamine-like substance and it may be significant in this connection that histamine has a powerful constricting action on the bronchioles in some animals. Relief is rapidly obtained on administration of adrenaline, isoprenaline and sometimes of antihistamines, when they bring about relaxation of the bronchial muscles.

Respiratory Movements

A constant renewal of air in the lungs is brought about by movements of the thorax, and this constitutes normal breathing. With inspiration the cavity of the thorax is enlarged, and the lungs enlarge as well to fill the increased space. As a result the capacity of the air passages of the lungs is increased, and air is drawn in through the trachea. Inspiration is immediately followed by expiration, which causes a diminution of the capacity of the thorax and expulsion of the air. During quiet breathing, expiration lasts 1·3 to 1·4 times as long as inspiration. Respiratory movements are to some extent under the control of the will, and one can, for instance, cease breathing for a time. It is impossible, however, to prolong this respiratory standstill for much more than a minute, for the urge to breathe becomes excessive and against our will we are forced to breathe.

Inspiration. This is achieved, first, by descent of the diaphragm which enlarges the thorax from above downwards. The diaphragm, which is innervated by the phrenic nerves, is the most important muscle of respiration although its exact contribution in normal quiet breathing varies widely in different subjects. During deep breathing it may be responsible for as much as 65 per cent. of the total volume of air inspired. Secondly, the thorax is enlarged by elevation of the ribs, mainly the second to the tenth. Since each pair of corresponding ribs forms a ring directed obliquely from behind downwards and forwards, this movement of the anterior end causes elevation of the sternum and an increase in the antero-posterior diameter of the thorax. At the same time there is an outward or lateral movement of each rib which increases the transverse diameter of the thorax.

The action of the intercostal muscles is at present uncertain but it would appear that in man at least, they are responsible for elevation of the ribs. When surface electrodes are placed over the intercostal muscles in the lower rib spaces, bursts of impulses occur only during inspiration in normal quiet breathing. Apparently there is no reciprocal action of the two layers of intercostal muscles.

Expiration. The muscles relax, the thoracic cage through its own elasticity contracts and the walls of the chest are brought closer together assisted by the recoil of the bronchial tree and the lung tissues (see Fig. 4. 2). In quiet breathing, expiration is probably entirely passive.

If a subject is observed in the supine position, it will be noticed that the abdominal wall rises and falls in each respiratory cycle. This is due to the movements of the diaphragm displacing the abdominal contents, and is sometimes referred to as "abdominal respiration."

The *abdominal muscles* form a group of muscles having two important mechanical actions : first, raising the intra-abdominal pressure which results in the abdominal contents being pressed against the diaphragm and force it to ascend ; and, secondly, drawing the lower ribs downwards and medially. These muscles show no electrical activity during quiet breathing. Activity appears, however, during the expiratory phase of

respiration when breathing is increased, and in other circumstances involving voluntary expiratory manœuvres, *e.g.* coughing.

The *accessory muscles of respiration* comprise the scalene and sternomastoid muscles, the pectorals and serratus anterior, the rhomboids, trapezius and latissimus dorsi. For the ribs to be raised effectively during inspiration by the intercostal muscles, the thoracic inlet, *i.e.* the first rib, must be fixed. This is done by contraction of the scalene muscles which show electrical activity only during inspiration. Of the other accessory muscles of respiration, the sternomastoid muscles are probably the most important and show activity during voluntary inspiratory efforts and in patients with dyspnœa, that is, difficulty in breathing.

The Pleura and the Pleural Cavities

The lungs are enveloped in a closed membranous sac. The outer wall of the sac lines the chest wall and is called the *parietal* pleura. This is reflected at the roots of the lungs, where it is continuous with the visceral layer which covers the lungs. The potential space between the layers is spoken of as the pleural cavity. It contains a small amount of fluid which acts as a lubricant. When the chest wall expands and contracts during the phases of respiration, the parietal and the visceral layers of the pleura normally maintain contact with each other. Even at the end of normal expiration the healthy adult lungs are in a stretched condition and this may be shown by the fact that if an opening is made in the pleural cavity, air rushes into the opening and the lungs collapse. The condition where air is present in the pleural cavity is called a *pneumothorax.* Since the lungs are always tending to collapse, it is evident that they must exert a pull on the thoracic wall. The pressure in the pleural cavity is called the *intrapleural pressure.* It may be measured either by injecting a very small volume of air between the two layers of pleura and inserting a needle, connected to a suitable manometer, into this pocket of air or, more conveniently, by measuring the intra-œsophageal pressure, a tube being passed through the mouth for this purpose. The walls of the œsophagus are lax and so the pressure in its lumen is practically the same as the intrapleural pressure.

Owing to the pull of the lungs, the intrapleural pressure is below the pressure of the surrounding atmosphere and is called a " negative pressure " in consequence. In the expiratory position, it is 3–5 mm. Hg below atmospheric pressure ; in normal inspiration it becomes 5–10 mm. Hg below atmospheric. If the lungs are more fully distended by a deep inspiration, the elastic forces are brought more into play, and the negative intrapleural pressure may amount to 30 mm. Hg.

In forced expiration with the glottis closed (Valsalva'a manœuvre) and in other circumstances involving expiratory efforts such as straining, coughing and defaecation, enormous increases in intrapleural pressure occur. The maximum pressure that can be maintained voluntarily for 1–2 seconds is about 110 mm. Hg. Transient intrapleural pressures of 300 mm. Hg may occur during severe coughing.

The lungs may be thought of as a pair of bellows, which are normally kept empty of air by a spring (Fig. 4. 3). In order to expand the bellows, we must pull on the handles with a force which is sufficient : (1) to extend the spring ; (2) to deform the material out of which the bellows are made ; and (3) to overcome the resistance to airflow through the nozzle. The first part of the force depends on the amount to which the spring is stretched—*i.e.* on the volume of air in the bellows. The two other parts depend on the *rate* at which the bellows are being expanded—*i.e.* on the rate at which air is being drawn in. Similarly, when the lungs are expanded during inspiration, the respiratory muscles must overcome both an *elastic resistance,* analogous to that exerted by the spring, and a *non-elastic* or " *viscous* " *resistance,* exerted by the lung tissue and by the flow of air through the bronchial tree.

If the lungs were perfectly elastic, and the viscous resistance negligible, the change in intrapleural pressure (analogous to the force pulling on the handle of the bellows) would depend only on the change in the volume of the lungs ; the pressure–volume relation would be the same

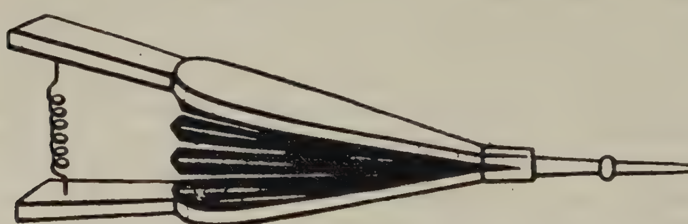


FIG. 4. 3. A simple model of the lungs represented by a pair of bellows and a spring. (From J. L. D'Silva, after R. V. Christie.)

whether the lungs were being expanded or allowed to collapse. Furthermore, during respiration, the greatest negative pressure would coincide with the maximum volume at the end of inspiration, and the least negative pressure with the minimum volume at the end of expiration ; the cyclic changes in intrapleural pressure and in tidal air volume would be in phase with each other. Examination of the records in Fig. 4. 4, however, shows that the two curves are not in phase ; the changes in intrapleural pressure precede the changes in tidal air volume. This is because of the viscous resistance of the lungs. At the beginning of inspiration, the intrapleural pressure falls more rapidly than would be expected, since not only has it to overcome the elastic resistance of the lungs, but also the viscous resistance, which increases as the rate of expansion of the lungs increases. In the same way, we have to exert an additional force on the handle of the bellows in order to deform the material and draw air in through the nozzle. Conversely, when the rate of expansion becomes smaller, towards the end of inspiration, this additional fall of intrapleural pressure also becomes smaller until, at the end of inspiration (and again at the end of expiration), the lungs are momentarily at rest and the " viscous " forces vanish. During expiration,

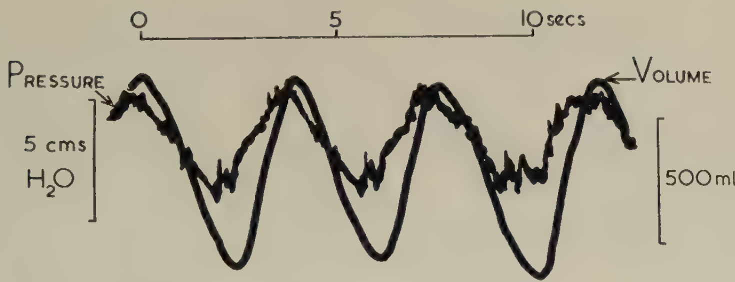


FIG. 4. 4. Simultaneous records of intra-œsophageal pressure (representing intratracheal pressure) and tidal air volume in a normal subject at rest. (McIlroy, Marshall & Christie.)

the rate of change of volume is reversed (the volume is decreasing, instead of increasing), the additional pressure necessary to overcome the viscous resistance is also reversed, and is subtracted from the pressure necessary to overcome the elastic resistance, instead of added to it ; the pressure–volume relation is a closed loop, as shown in Fig. 4. 5.

The existence of viscous, as well as elastic, resistance has two consequences of some importance. First, it contributes to the work

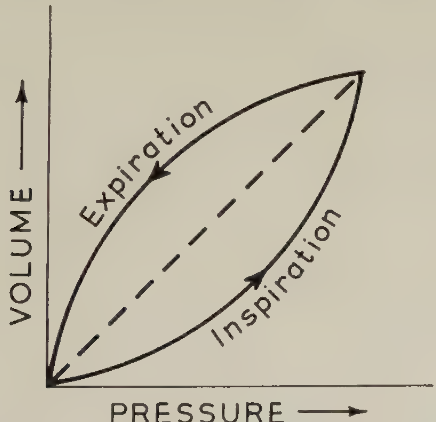


FIG. 4. 5. Diagram showing the relationship between intrapleural pressure and tidal volume. The interrupted line shows the relationship were the lungs to comprise only an elastic resistance. The continuous line shows the relationship during inspiration and expiration with the viscous component added.

which must be done and to the oxygen consumed by the respiratory muscles in order to ventilate the lungs : this oxygen consumption is added to that needed for more " useful " work and of itself necessitates an increase in the volume of air inspired. The work done in overcoming the viscous resistance is about 40 per cent. of the total, which increases disproportionately as the respiratory minute volume increases. Normal values are 0·5 ml. oxygen per litre of ventilation, or 0·5 kg-m/min

at rest ; when a subject breathes maximally (about 180 litres/min) the work of breathing amounts to 250 kg-m/min. Secondly, the ratio of the viscous resistance to the elastic resistance determines the speed at which the lungs collapse. If, for some reason, the viscous resistance increases without a corresponding increase in the elastic resistance (the strength of the spring in Fig. 4. 3), the lungs may not return to the initial expiratory state before the next inspiration begins. The volume of air in the lungs will become progressively greater (the spring will be increasingly stretched) until a new steady state is reached and the work needed to maintain the respiration will be greatly increased. This may occur during an attack of asthma.

Fœtal Respiration and Expansion of the Lungs at Birth

The lungs of the fœtus are airless and unexpanded ; they contain a small amount of amniotic fluid. Only about 20 per cent. of the right ventricular output passes through the pulmonary circulation, the greater part of the blood flow being short-circuited by an anastomotic channel between the pulmonary artery and aorta, the ductus arteriosus. It is probable that small but ineffective respiratory movements are made *in utero*. In animals these movements become accentuated when the mother is given a gas mixture containing little oxygen to breathe, and sometimes when the carbon dioxide content is increased. Increased respiratory movements also occur when the placental circulation is obstructed.

At birth, respiration is brought about by contraction of the diaphragm and intercostal muscles. At the same time, the pulmonary circulation is established although it is some hours before the ductus arteriosus closes completely. There are probably many factors coming into play in promoting the first breath. Of major importance in this connexion are the impulses emanating from the skin and from proprioceptors in the muscles, tendons and joints after birth of the fœtus into the atmospheric air. Not only is the fœtus exposed to the cold air and to contact with surrounding objects, but its transfer from its aqueous environment subjects it to the strains put upon it by its own weight. Another factor which may be important in bringing about the first breath is the chemical change in the blood when the umbilical cord is tied or the placental circulation fails ; the oxygen supply to the fœtus is cut off and carbon dioxide produced by the body accumulates. Consequently a state of asphyxia results in which the arterial blood oxygen pressure falls and carbon dioxide pressure rises. But how these changes affect respiration in the fœtus is not fully understood.

The pressure necessary to inflate the fœtal lungs for the first time is much greater than for subsequent breaths. This is due to the fact that in the fœtal lung, two resistances have to be overcome : (1) that due to the cohering bronchiolar and alveolar surfaces which have to be separated by the residual volume of air, and (2) the resistance offered by the elastic tissue and smooth muscle of the lung parenchyma. The

first of these is considerable and as a result pressures of up to -35 mm. Hg must be applied to the outside of the fœtal lungs to inflate them.

Sounds Associated with the Movements of the Lungs

There are two distinct sounds associated with the movement of air into and out of the normal lungs, both of which may be heard by placing the ear on the chest or, better, by using a stethoscope. The first of these is a fine rustling noise, occurring during inspiration and the beginning of expiration, known as *vesicular breathing* ; the second, heard only when the stethoscope is placed over one of the larger air passages, is louder and rougher, like a whispered " hah," and is known as *bronchial breathing*.

The Primary Subdivisions of the Lung Volume

The diagrammatic representation of a tracing of respiratory movements shown in Fig. 4. 6 was taken from a subject who, after a few

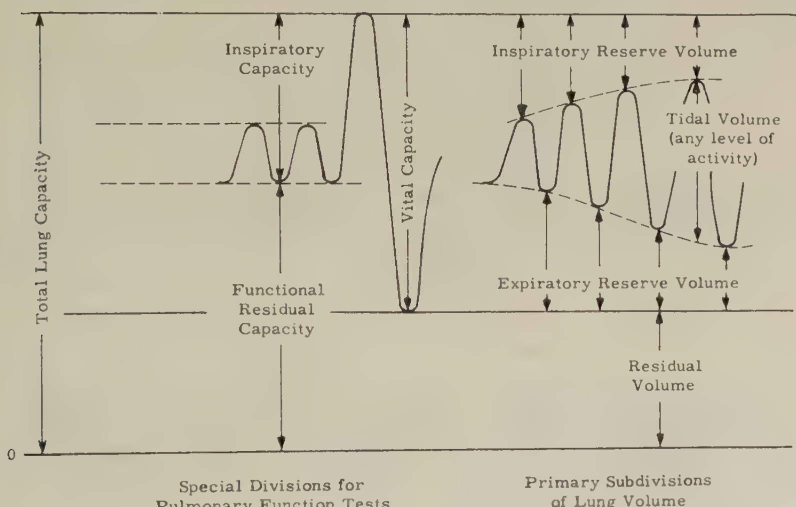


FIG. 4. 6. Diagram showing lung volumes and capacities.

normal quiet respirations, took the deepest inspiration that he could and then expired to the limits of his ability. The upward stroke represents inspiratory and the downward stroke expiratory movements. The lung volumes may be divided into two main divisions, the " volumes " which can be breathed by the subjects, these being dynamic volumes, and the " capacities," each of which includes two or more of the primary volumes.

Values for the volumes are obtained by breathing in and out of a *spirometer* such as that shown in Fig. 4. 7. This is a volume recorder on the lines of a gasometer ; a cylindrical bell, closed at the upper end, is immersed in a tank of water, and counterbalanced. Air blown through a pipe passing through the water-seal raises the bell and the distance the bell travels is recorded by a writing-point on a moving paper. The spirometer is calibrated by injecting into it known volumes of air.

The volume that is inspired and expired during quiet breathing is referred to as the **tidal volume** and varies from **350** to **600** ml. The tidal volume multiplied by the frequency of the respirations per minute gives the total volume of air breathed per minute ; this is known as the **respiratory minute volume.** At rest this is usually 4 to 8 litres per minute. The volume of the deepest inhalation that can be taken at the end of a normal inspiration is called the **inspiratory reserve volume** and is about **2,500** ml. The deepest exhalation possible at the end of a normal expiration is called the **expiratory reserve volume** and is about **1,500** ml.

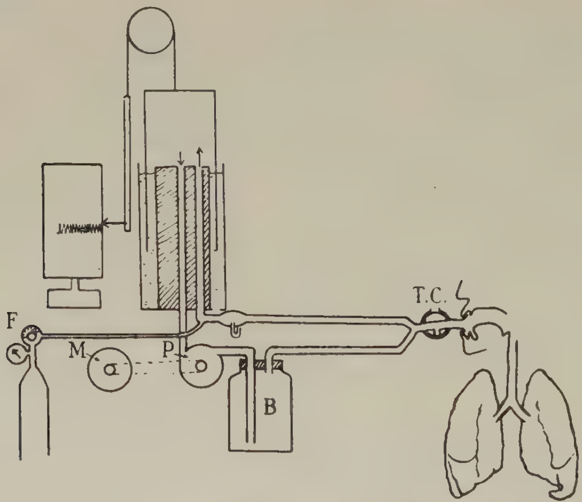


FIG. 4. 7. Spirometer arranged in a closed circuit respiratory system to determine lung volumes. T.C., tap ; B, bottle with absorbent for CO_2 (soda-lime or caustic potash solution) ; P, air-circulating pump driven by motor, M ; volume of the system maintained constant by running in 0_2 from cylinder F at the same rate as usage. (Redrawn from Herald and McMichael.)

Vital Capacity. When a subject, after the deepest inspiration, expires the largest volume that he can into a spirometer, the volume of air that he expires is termed his *vital capacity*. Reference to the diagram in Fig. 4. 6 shows that this volume is the sum of his tidal, inspiratory and expiratory reserve volumes and amounts to 4–5 litres. This amount is not an expression of the total volume of air in his lungs, since when he has completed his forced expiration there still remain about 1,500 ml. of the *residual volume* which could only be expelled by opening the chest wall and allowing the lungs to collapse. To determine the total capacity of the lungs this residual volume must be estimated and added to the vital capacity. Even in the collapsed lung a small volume of the air is trapped ; it is this which gives the collapsed lungs their buoyancy in water and is used in medico-legal investigations to ascertain whether the lungs have ever expanded and so determine whether a child breathed after birth.

In the past, vital capacity has been taken as a measure of respiratory function, but it is now thought that a single measurement gives little useful information, whereas a change in the vital capacity may be significant. Thus, the vital capacity of a patient suffering from active lung disease falls as the disease progresses.

Probably a more useful test of respiratory function is the maximal breathing capacity (M.B.C.). The subject ventilates maximally for a period of fifteen seconds and the value obtained for the respiratory minute volume is then expressed in litres per minute. Normal values range from 80 to 170 litres/min.

The vital capacity is affected by the volume of blood in the pulmonary circulation. It is smaller when the subject is supine than it is when he is standing, the larger volume of blood in the pulmonary vessels presumably encroaching on the air capacity of the lungs. In certain forms of heart failure, also, particularly in left ventricular failure where the pulmonary vessels are congested, there is a reduction in vital capacity. These conditions are often accompanied by respiratory distress or difficulty in breathing (dyspnœa), but the mechanism by which it is brought about is still uncertain.

The *inspiratory capacity* is the maximum volume of gas that can be inspired from the resting expiratory level and is the sum of the inspiratory reserve volume and the tidal volume (Fig. 4. 6).

The volume of gas remaining in the lungs at the resting expiratory level is known as the *functional residual capacity*. Reference to Fig. 4. 6. will show that this is the sum of the residual volume of the lungs and the expiratory reserve volume.

Composition of the Respired Air

Expired Air. The tissues of the body use oxygen for the oxidation of various materials and in consequence produce carbon dioxide. A man weighing 70 kg. consumes about 250 ml. oxygen/min and produces about 200 ml. carbon dioxide/min. The blood reaching the lungs contains more carbon dioxide and less oxygen than arterial blood and in passing through the lungs therefore gives off carbon dioxide and takes up oxygen through an interchange with the air in the alveoli. This air is continually renewed by breathing, and hence expired air contains less oxygen and more carbon dioxide than that which is inspired. The composition of expired air, however, is not constant.

A sample of expired air is obtained by making the subject breathe through valves so arranged that he breathes in from the atmosphere and out into an airtight bag, known as a *Douglas* bag, from the name of its first user. Samples of air for analysis are drawn off from the bag and the volume of air collected in the bag in a given time is measured by pressing out its contents through a gas meter.

The expired air of a normal resting subject at sea-level contains 3·0–4·5 per cent. of carbon dioxide and 16·0–17·5 per cent. of oxygen. A comparison of the composition of inspired (atmospheric) and expired air is shown in Table 4. 1.

As mentioned above, the volume of oxygen used up per minute

is larger than the volume of carbon dioxide added to the expired air. The total *volume* of nitrogen leaving the lungs is, however, the same as that taken in. In the above analyses, the nitrogen percentage is seen

TABLE 4. 1

Composition of inspired and expired air

	Inspired air (per cent.)	Expired air (per cent.)
Oxygen	20·95	16·4
Carbon dioxide	0·04	4·1
Nitrogen (including argon)	79·01	79·5

to be higher in the expired air. This is because the volume at N.T.P. of air expired is less than that inspired, owing to the disappearance of a certain amount of oxygen without the production of a corresponding amount of carbon dioxide, so that the relative amount of nitrogen is slightly increased.

Alveolar Air. Of the 500 ml. of air drawn in during an average breath, only about 350 ml. reach the alveoli. The other 150 ml. are contained in the conducting air-way from the nose down to the respiratory bronchioles. This volume is known as the *respiratory dead space* and does not participate in gas exchange in the lungs. Hence alveolar air must contain more carbon dioxide and less oxygen than mixed expired air which consists of alveolar plus unchanged dead space air.

At the end of an expiration the dead space has been swept out by, and remains filled with, alveolar air. Based on this fact, a sample of alveolar air may be obtained for analysis in the following way. A piece of india-rubber tubing, 25 mm. diameter and 1·5 m. long, is fitted with a mouth-piece, near to which is connected a gas-sampling tube which is provided with a three-way tap at each end. Before an experiment, the sampling tube is evacuated. The subject of the experiment applies a nose-clip and after breathing normally a few times, puts his mouth to the tube at the end of a normal inspiration, expires quickly and deeply and closes the mouth-piece with his tongue. The tap of the sampling-tube is then turned, and the air near the mouth-piece, which is that last expelled from the lungs, rushes into it. The tap of the tube is then turned off, and the gas sample removed for analysis. A similar sample is then taken, in which the subject expires deeply at the end of a normal expiration. This sample will contain slightly more CO_2 and less O_2 than that obtained at the end of inspiration. The mean of the two samples is taken as the average composition of the subject's alveolar air.

The significance of the composition of alveolar air will be better understood by reference to Fig. 4. 8. This is a diagrammatic representation of what happens to the pressure of the carbon dioxide and oxygen in the alveolar air and in the blood as it circulates through the lungs. Let us suppose that a sample of the venous blood entering the lungs by

the pulmonary artery contains carbon dioxide at a pressure of 46 mm. Hg. As this sample circulates through the pulmonary capillaries some of the carbon dioxide will diffuse out into the air in the alveoli and will continue to diffuse until there is no further pressure difference between blood and alveoli. If the pressure of carbon dioxide in the alveolar air is 40 mm. Hg, the pressure of carbon dioxide in the arterial blood in the lung circulation will also reach this value. This means that the carbon dioxide pressure in the alveolar air is the same as that of the arterial blood leaving the lungs. The evidence that such an equilibrium exists will be considered in a later section.

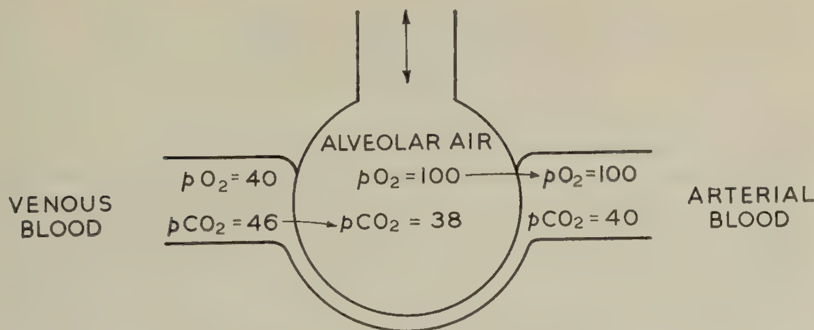


FIG. 4. 8. Diagrammatic representation of alveolar air (see text).

The average composition in volumes per cent. of alveolar air for adult men at rest, at 760 mm. Hg pressure, is shown in Table 4. 2, together with the partial pressures exerted by the constituent gases. It is often better to express the data regarding respiratory gases in terms of partial pressures, as in the last column of this table, than in terms of percentage composition.

TABLE 4. 2

Composition of alveolar air

	Volumes (per cent.) (dry)	Partial Pressures (mm. Hg) (wet)
Carbon dioxide	5·6	40
Oxygen	13·8	99
Nitrogen	80·6	574
Water vapour	0·0	47
	100·0	760

From Dalton's law we know that the total pressure exerted by a mixture of gases is equal to the sum of the separate pressures which each gas would exert if it alone occupied the whole volume. Alveolar air is in contact with wet tissues and is saturated with aqueous vapour ; its

partial pressure, depending only on the temperature (47 mm. Hg at 37° C.), must be included as a part of the total (*i.e.* barometric) pressure. The following example will explain the calculation.

Suppose that a sample of alveolar air had the composition (dry), $CO_2 = 5{\cdot}6$ per cent.; $O_2 = 13{\cdot}7$ per cent.; $N_2 = 80{\cdot}7$ per cent. by volume, and that the barometric pressure was 755 mm. Hg at the time. Then, since the aqueous vapour pressure = 47 mm. Hg, the remaining gases together contribute (755 − 47) = 708 mm. Hg, and of this

$$CO_2 = \frac{5{\cdot}6}{100} \times 708 = 40 \text{ mm. Hg}$$

$$O_2 = \frac{13{\cdot}7}{100} \times 708 = 97 \text{ mm. Hg}$$

$$N_2 = \frac{80{\cdot}7}{100} \times 708 = 571 \text{ mm. Hg.}$$

These are the pressures exerted by each gas on any boundary surface.

Dead Space. As has already been mentioned, this refers to those parts of the respiratory tissues—the nose, pharynx, trachea, bronchi and bronchioles—which act as a conduit for the passage of gases to the alveoli. This is known as the *anatomical dead space*. The *physiological dead space* includes the anatomical dead space and two additional volumes : (1) the volume of inspired gas which ventilates alveoli which receive no pulmonary capillary blood flow, and (2) the volume of inspired gas which ventilates alveoli in excess of that volume required to arterialise the blood in pulmonary capillaries. In the normal subject, these two additional volumes are negligible and so the anatomical and physiological dead spaces are equal.

The volume of the dead space cannot be measured directly, but can be calculated using Bohr's equation. This states simply that the gas expired from the lungs is a mixture of gas from the dead space and from the alveoli ; if two of these (the expired air and alveolar air) are known, the third (the dead space air) can be calculated. From Tables 4. 1 and 4. 2, we can take the following representative figures : (1) CO_2 in alveolar air : 5·6 per cent. ; (2) CO_2 in expired air : 4·1 per cent. We will take the volume of tidal air as 600 ml. (all measured dry and at room temperature). Then the total quantity of carbon dioxide expired in each breath is $\frac{600 \times 4{\cdot}1}{100} = 24{\cdot}6$ ml. But this quantity of carbon dioxide is contained in $24{\cdot}6 \times \frac{100}{5{\cdot}6} = 440$ ml. of alveolar air ; so that the 600 ml. of expired air has the same composition as 440 ml. of alveolar air mixed with 160 ml. of inspired air, containing no carbon dioxide. The dead space consequently would have a volume of 160 ml. It must be remembered that this is the volume of the gas in the dead space measured dry and at a lower temperature than that at which it was when in the body. But since we wish to measure the volume of a cavity within the body, we must correct this gas volume to body

temperature (37° C.) and saturate with water vapour (47 mm. Hg). (For details see textbooks of practical human physiology.)

The volume of the dead space varies, not only in different individuals, but even in the same individual according to posture and other factors. The most reliable measurements indicate that normally the dead space volume varies between 100 and 250 ml.

In disease, the physiological dead space may greatly exceed the anatomical dead space when ventilation in parts of the lungs is greater than that required to arterialise the blood.

Alveolar ventilation. We have seen how the product of the tidal volume and frequency of breathing gives a value for the respiratory minute volume, or the turn-over of air as measured at the mouth. It must be realised, however, that part of the inspired air at each breath never reaches the alveoli, but remains in the dead space and is in consequence expired again without undergoing any change. The respiratory minute volume therefore gives little indication of the turn-over of air in the *alveoli*, which is of vital importance when considering how effectively oxygen is supplied to the alveoli and carbon dioxide is got rid of during breathing. The turn-over of air in the alveoli is known as the *alveolar ventilation* and is the product of the (tidal volume — dead space) and frequency of breathing. To illustrate how changes in breathing may affect the respiratory minute volume and alveolar ventilation differently we may consider the following example. A subject has a tidal volume of 500 ml., respiratory frequency 16/min and dead space 150 ml. His respiratory minute volume $= 500 \times 16 = 8{,}000$ ml./min. On the other hand, his alveolar ventilation $= (500 - 150) \times 16 = 350 \times 16 = 5{,}600$ ml./min. Now, if the subject doubled his tidal volume (to 1,000 ml.) and halved his frequency (to 8/min), his respiratory minute volume would remain unchanged. His alveolar ventilation, however, would increase to 6,800 ml./min. This emphasises the importance of the measurement of the volume of inspired air entering the alveoli rather than that entering the upper respiratory tract as a measure of the turn-over of air in those parts of the lungs concerned with gas exchange.

The Mechanism of Gaseous Exchange between the Lungs and the Blood

In the lungs, carbon dioxide leaves the blood stream and enters the air in the lung alveoli, and oxygen leaves the alveoli and enters the blood. Since in doing so the gases have to pass through two membranes, the walls of the alveoli and the walls of the capillaries, the question arises as to whether the interchange of gases between the alveolar air and the blood in the pulmonary capillaries can occur by the purely physical process of diffusion. To answer this question we must study the partial pressures or tensions of oxygen and of carbon dioxide in the alveolar air, in venous blood coming to the lungs and in the arterial blood leaving the lungs. If the process is one of diffusion, the partial pressures and hence the flow of oxygen must be, in descending order :

alveolar air, arterial blood, venous blood ; and those of carbon dioxide in the reverse order.

The partial pressure of oxygen in the arterial blood may be estimated in three ways, the principles of which are briefly as follows.

(1) The *Aerotonometer*. A very small volume of air is brought into equilibrium with a relatively large volume of blood, which has been withdrawn from the subject and made incoagulable without being allowed to come into contact with the air. The small bubble of air is then analysed, and since the actual amount of gas that has been exchanged with the blood is very small, the partial pressures of the gases in the bubble are sensibly the same as those originally in the blood.

(2) The percentage saturation of the blood with oxygen is determined on one portion of a sample withdrawn from the subject, and an oxygen dissociation curve is constructed on another portion ; the partial pressure of oxygen corresponding to the percentage saturation of the first portion can then be read off. Care must be taken, of course, that the dissociation curve is made at the same partial pressure of carbon dioxide as that in the first portion of the sample.

(3) The subject is made to breathe a gas mixture containing known partial pressures of oxygen and of carbon monoxide. The ratio of the concentrations of oxyhæmoglobin and of carboxyhæmoglobin in the subject's arterial blood is determined either by the reversion spectroscope (Hartridge) or by comparison with mixtures of oxyhæmoglobin and carmine in dilute solution (Haldane). It is then assumed that no secretion of carbon monoxide can occur, and that the following relation holds in the blood :

$$\frac{[O_2Hb]}{[COHb]} = K.\frac{pO_2}{pCO}$$

The square brackets indicate concentrations. The value of the constant K can be determined *in vitro* : we know the value of pCO and the value of the ratio on the left side of the equation ; we can thus calculate the value of pO_2.

The results of determinations made on man under normal conditions are summarised in Fig. 4. 8 and show that the partial pressure of oxygen in the arterial blood is very close to that in the alveolar air. The pressure of oxygen in the alveoli is about 100 mm. Hg, and the venous blood, therefore, with an oxygen pressure of 40 mm. Hg, brought to the alveoli will rapidly take up oxygen from them and approach the point of saturation.

The pressure of carbon dioxide in venous blood is about 46 mm. Hg and that of the alveolar air is about 40 mm. Hg. This pressure gradient will tend to cause a flow of carbon dioxide from the blood into the alveoli.

It is now generally agreed that the respiratory gas exchange is accomplished by a process of simple diffusion, the direction and extent of which depends almost entirely upon the difference in tension or pressure on the two sides of the alveolar membrane, that is, the molecules of oxygen pass from a region of high partial pressure to one of lower partial pressure.

An alternative view that the pulmonary epithelium actively secretes oxygen was put forward by J. S. Haldane and his co-workers based on observations made on the summit of Pike's Peak (14,000 ft.) in 1912. The results indicated that the arterial oxygen pressure (as measured by the carbon monoxide method and carmine titration) could exceed the alveolar oxygen pressure, so that secretion must have been taking place. Later work of

Barcroft and his co-workers at Cambridge (1921) in an atmosphere corresponding to that at a height of 18,000 ft., and using arterial puncture and an aerotonometer, failed to confirm this ; there was no need to presume secretion. Barcroft's views were later confirmed by Dill and his collaborators, the average of a series of analyses in ten subjects at 17,500 ft. showing no significant difference between the pressure of the oxygen in the lungs and that in the arterial blood. In the mountains of the Andes, moreover, there are natives who for many generations have lived all their lives at a high altitude, but whose arterial blood has a low saturation with oxygen. Even they, apparently, have not been able to secrete oxygen in their lungs very effectively, if at all.

Diffusion. In the lungs of man, about 3 litres of alveolar air, corresponding to the functional residual capacity (see Fig. 4. 6), surround about 100 ml. blood contained in the capillary network of the pulmonary vascular bed. The pulmonary capillaries provide an area of about 40 m^2, and a freely permeable membrane less than $1{\cdot}5\mu$ thick, separating the blood from the alveolar air. The volume of gas which is capable of passing across the alveolar membrane is known as the *diffusing capacity of the lungs.* It is defined as the quantity of gas transferred each minute for each millimeter of Hg difference in partial pressure of the gas on the two sides of the membrane. Apart from being dependent on the difference in partial pressure of the gas in the alveolar air and in the pulmonary capillary blood, the diffusing capacity is (1) proportional to the total area available for diffusion, (2) inversely proportional to the average thickness of the alveolar membrane, (3) proportional to the ease with which the gas diffuses through the type of tissue comprising the alveolar membrane, and (4) proportional to the solubility of the gas in the alveolar membrane. The solubility of carbon dioxide in watery solutions (and tissues) is very much greater than that of oxygen, and this explains why almost the same volume of carbon dioxide as oxygen diffuses across the pulmonary membrane in spite of the fact that the gradient for carbon dioxide is only 6 mm. Hg (46 — 40), as against 60 mm. Hg (100 — 40) for oxygen.

The normal value for diffusing capacity of normal young adults is 20–30 ml. oxygen/min for each mm. Hg partial pressure difference. It more than doubles during exercise on a treadmill due to an increase in surface area for diffusion caused by opening up of additional pulmonary capillaries and to dilatation of others already patent. The maximum diffusing capacity decreases with age in the later decades of life, probably due to a reduction in the number of pulmonary capillaries.

Methods of Recording the Respiration. Ideally, the record of the respiratory movements should allow us to discover (*a*) the frequency of respiration, (*b*) the depth of respiration and (*c*) the degree of expansion of the lungs at any moment, even if respiratory movements have stopped. There are only two ways in which this can be done. The first, which is shown in Fig. 4. 7, is to connect the mouth of the subject (or the tracheal cannula of an animal) with a spirometer, to remove continuously the carbon dioxide from the expired air, and to add oxygen to make up for that used in the metabolism—to make use, in fact, of exactly the same apparatus as is used for determining the metabolic rate, as will be described in Chapter 6. The other way is to enclose the whole body, with the exception of the head, in an airtight box ; the hole through which the head protrudes being sealed by a rubber

collar around the neck. A spirometer connected with the box then records the respirations completely. An " iron lung " (p. 143) could be used for this purpose.

The simpler forms of apparatus for recording the respiration in man do not give accurate measurements of the total ventilation, but are useful, nevertheless, for many purposes. The *stethograph*, as now commonly used, consists of a piece of large diameter rubber tubing stoppered at both ends and connected with a tambour by a side tube at the centre ; this is tied round the chest or abdomen, the movements of which distort it, and drive air into or out of the tambour. Many other devices for recording the movements of the chest or abdomen have also been described.

The Regulation of Breathing

Discharge in efferent nerves. The regulation of breathing is brought about by alterations of the activity of the groups of motor neurones which control the respiratory muscles. The dominant motor neurone pools are collectively known as the *respiratory centres* and these are controlled through agents acting directly on the centres and also reflexly through the arrival of afferent impulses.

The main muscles of respiration are the diaphragm, innervated by the phrenic nerves which originate from the spinal roots C_2, C_3 and C_4, and the intercostal muscles which receive their innervation via the intercostal nerves, T_1–T_6. Each act of inspiration involves a discharge along these and a number of other nerves, *e.g.* the facial to the muscles moving the alæ nasi and the vagus to the muscles of the larynx. Thus many segmental levels are involved in the innervation of the muscles of respiration, and the respiratory act is probably therefore integrated in the brain stem rather than in the spinal cord.

Studies of the activity of motor units in the diaphragm and intercostal muscles have shown that there is activity throughout the whole phase of respiration, but that it is greater in inspiration than in expiration. During expiration the rate of firing is slow and is responsible for the thorax being maintained in a state of partial inspiration. The periodic increases and decreases in volume of the thorax are therefore superimposed on an underlying postural tone. The expiratory muscles also exhibit tonic activity. The act of inspiration begins, therefore, on a background of tonic activity in both the inspiratory and expiratory muscles, but whereas the activity of those neurones supplying inspiratory muscles which are in tonic contraction increase their rate of firing, that of expiratory neurones is reciprocally inhibited. Furthermore, as inspiration proceeds, new motor units come into action, or are " recruited," as shown in Fig. 4. 10, so that the inspiratory act gains force as it proceeds. Then, when inspiration reaches a peak, it is abruptly terminated by various factors which control the depth of respiration. As expiration occurs, the units which maintain inspiratory tone revert to their former steady rate of firing, whereas the tonic discharge in expiratory units returns.

Respiratory Centres. Section of the brain stem above the level of the pons in the dog and cat does not affect breathing, but respiratory activity ceases altogether when all but the lower third of the medulla

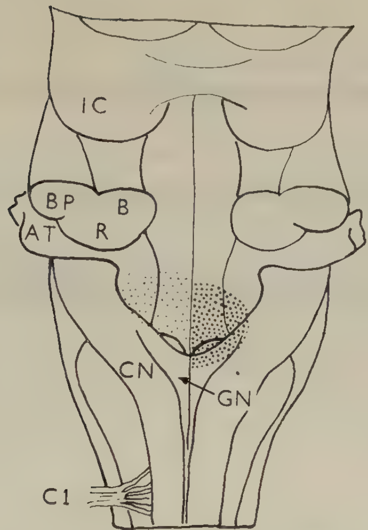


Fig. 4. 9. Dorsal view of the brain stem (cat) with cerebellum removed, showing the projection of the medullary respiratory centres on the floor of the fourth ventricle. To avoid overlapping, the *expiratory* centre is shown only on the left, and the *inspiratory* centre only on the right. IC, inferior colliculus ; BP, brachium pontis (middle cerebellar peduncle) ; B, brachium conjunctivum (superior peduncle) ; R, restiform body (inferior peduncle) ; AT, acoustic tubercle ; CN, cuneate nucleus ; GN, gracile nucleus ; C1, first cervical root. (Pitts, Magoun & Ranson, 1939.)

is removed. This indicates that the centres controlling respiration lie in the pons and upper part of the medulla. They are subdivided into the *inspiratory* and *expiratory* centres in the medulla, and a *pneumotaxic* centre in the pons. The inspiratory centre is sensitive to changes in the chemical composition of the blood and, in particular, to the partial pressure of carbon dioxide (pCO_2) and hydrogen ion concentration, and to a less extent to the partial pressure of oxygen (pO_2).

Our present knowledge as to the location of these centres is largely the result of the work of Pitts and his associates. They used concentric needle electrodes to stimulate various parts of the brain stem and found areas in the medulla which when excited caused an inspiratory or expiratory response corresponding to the centres mentioned above. Comroe showed, moreover, that localised injections of a carbon dioxide-bicarbonate buffer solution at suitable points led to a normal hyperpnœa, though he was unable to produce maintained inspiration in this way.

The *inspiratory centre* is in the ventral reticular formation immediately over the cephalic four-fifths of the inferior olive at the level of the entrance of the vagus nerve (Fig. 4. 9). Electrical stimulation of this region causes deep inspiration involving both the thorax and diaphragm.

The *expiratory centre* consists of neurones which intermingle with those of the inspiratory centre but extend slightly higher up in the

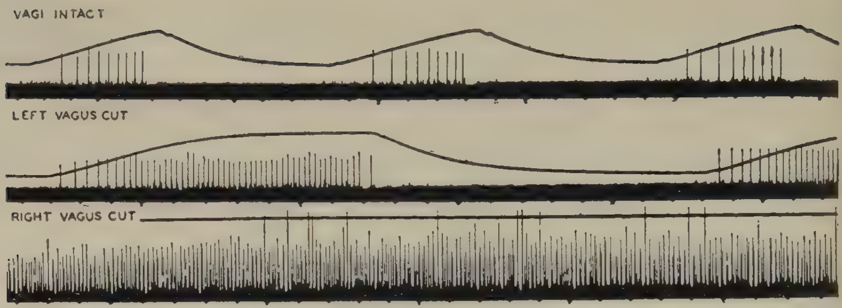


FIG. 4. 10. The effect of cutting the vagus nerves on respiratory activity in a cat after section of the brain stem immediately below the pneumotaxic centre. The lowest record shows the production of apneusis.
Each record shows, from above downwards: respiratory movements (upstroke = inspiration); nerve impulses in two phrenic motoneurones; time in 0·2 second. (Pitts, 1942.)

medulla and lie more dorsally and laterally in the reticular formation (Fig. 4. 9). Electrical stimulation of the expiratory centre causes cessation of respiration in the expiratory position and contraction of the expiratory muscles.

The *pneumotaxic centre* lies in the upper part of the pons and is responsible, in part, for periodically inhibiting the inspiratory centre and thereby converting its spontaneous continuous discharge into a rhythmic pattern of discharge, corresponding to inspiration, and inactivity, expiration. The circuit is: inspiratory centre → pneumotaxic centre → expiratory centre, which are excited in that order. As shown in Fig. 4. 11, when the state of excitation of the expiratory centre reaches a suitable level, impulses pass from it to the inspiratory centre, which is then inhibited, and this brings the inspiratory cycle to an end. This in turn causes the drive to the expiratory centre through the pneumotaxic centre, to be cut off, and the dominant inspiratory centre then emerges and starts off the next respiratory cycle.

Another way in which the discharge from the inspiratory centre is inhibited at the height of inspiration is by the arrival of vagal impulses from stretch receptors in the lungs. As the lungs inflate during inspiration, an increasing number of impulses per second reach the expiratory centre which in turn inhibits the inspiratory centre to cut short the inspiratory act. This vagal reflex is known as the Hering-Breuer reflex and is described more fully on p. 126. In so far as the normal control of breathing is concerned these vagal afferent impulses convert the steady discharge of impulses from the inspiratory centre into a rhythmic alternation of activity leading to inspiration and expiration. Their effect on the inspiratory centre is therefore similar to that of the impulses from the pneumotaxic centre, but is more powerful. This may be shown in the following way:

(1) If both vagus nerves are cut, so as to exclude afferent impulses from the lungs, breathing becomes slower and deeper (Figs. 4. 10 and

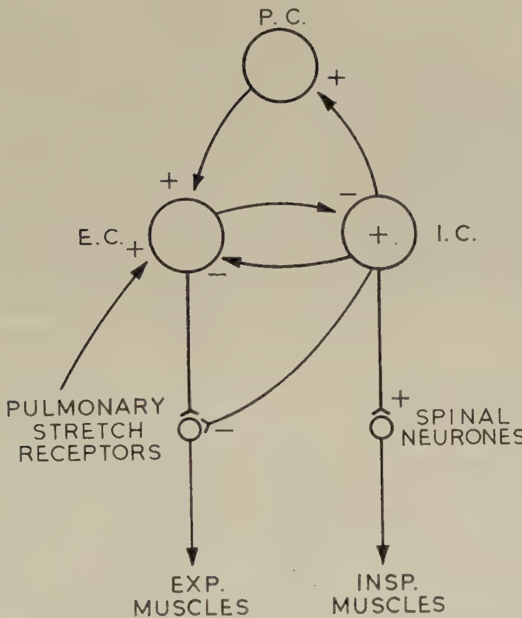


FIG. 4. 11. Diagrammatic representation of the respiratory centres. I.C., inspiratory centre ; E.C., expiratory centre ; P.C., pneumotaxic centre. Stimulus originating in the inspiratory neurones (+) spreads, as shown, to the expiratory centre, inhibiting it (−), and to the pneumotaxic centre. There is also direct reciprocal inhibition of the expiratory muscles (−). The discharge from the inspiratory centre is then inhibited by impulses from (1) pulmonary stretch receptors, via the vagus nerves, and (2) the pneumotaxic centre, both acting through the expiratory centre.

4. 12). On the other hand, if in an animal with intact vagi, a section through the upper pons is made so as to cut off the pneumotaxic centre, breathing continues normally.

(2) If the pneumotaxic centre is cut off and both vagi are divided so as to isolate the inspiratory centre, rhythmic breathing ceases and prolonged inspiratory spasms occur. This is known as *apneustic breathing* or *apneusis* (Figs. 4. 10 and 4. 12). If this activity of the inspiratory centre is studied by recording action potentials in a single fibre of the phrenic nerve, it can be shown that in apneusis there is a steady discharge of impulses at a high frequency in contrast to the waxing and waning of activity observed in normal breathing (Fig. 4. 10). Electrical stimulation of the central end of a cut vagus will arrest the inspiratory discharge, which returns when the stimulus is stopped.

From this it is clear that the inherent steady discharge of impulses from the inspiratory centre is converted into a rhythmic activity by two inhibitory mechanisms brought into action automatically but indirectly by the activity of the inspiratory centre itself, viz., the discharge from the pneumotaxic centre, and the discharge from pulmonary stretch receptors evoked by distension of the lungs.

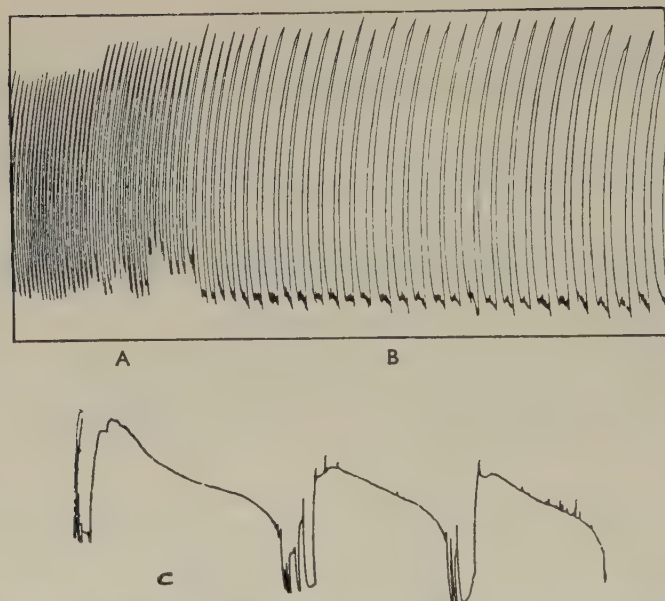


FIG. 4. 12. Respiratory movements of a cat, showing the **Different Types of Respiration.**

A. **Normal** respiratory movements (inspiration upwards).
B. The same after **Section of the Vagi.**
C. The **Apneustic** type, with prolonged inspiratory tonus.

The Chemical and Reflex Control of Breathing

By far the most important factor in the control of breathing is the sensitivity of the respiratory centre to changes in the carbon dioxide pressure of its blood supply. This is known as the *chemical* control. Normal breathing in the resting subject continues its even tenor so long as the carbon dioxide pressure in the alveolar air, and so in the arterial blood leaving the lungs, remains constant at about 40 mm. Hg. If the carbon dioxide pressure falls below this level breathing is inhibited, until carbon dioxide accumulation once more restores it to the normal value. On the other hand, a rise in the partial pressure of the alveolar carbon dioxide acts as a respiratory stimulant and the pulmonary ventilation is increased in an effort to maintain the alveolar partial pressure of carbon dioxide at the normal value, or nearly so. The part played by oxygen in the chemical control of breathing is, at sea-level, a relatively minor one.

In addition, the activity of the respiratory centre is affected by reflex mechanisms (Fig. 4. 21, p. 129). The afferent pathways for these are :—

(1) The carotid sinus nerves and the aortic (depressor) nerves carrying impulses from chemoreceptors situated in the bifurcation of the common carotid artery, and in the arch of the aorta, respectively.

(2) The vagus nerves carrying impulses from stretch receptors in the lungs.

(3) Afferent fibres from proprioceptors in the limbs.

(4) Afferent fibres from receptors of various kinds in the skin and the mucous membranes of the respiratory tract.

Chemical Control. The way in which the respiratory centres respond to carbon dioxide can be demonstrated by a number of simple experiments.

(1) **Asphyxia.** The proper aeration of the blood may be interfered with by allowing a subject to re-breathe repeatedly in and out of a small bag. If his respiratory movements are recorded, they will be seen to increase gradually in depth and frequency. This is known as *hyperpnœa.*

(2) **Hypercapnia.** In asphyxia, two factors are present, first an increase in the amount of carbon dioxide, and secondly a diminution in the amount of oxygen in the blood. Each factor has certain effects on breathing and it is important, therefore, to distinguish *oxygen lack* or *deficiency,* and excess of carbon dioxide or hypercapnia. These two states can be separated in a simple experiment. If a subject is made to re-breathe from a bag containing room air, respiration becomes noticeably increased when the concentration of carbon dioxide in the bag reaches about 3 per cent. and that of oxygen falls to about 17 per cent. If the experiment is repeated, but this time with a soda lime tower situated between the subject's mouth and the bag to absorb the carbon dioxide and prevent it accumulating in the bag, there is no appreciable increase in respiration until the oxygen concentration in the bag falls to about 14 per cent. (The carbon dioxide concentration in the bag is now, of course, zero.) Finally, if the subject is made to breathe 100 per cent. oxygen and the carbon dioxide is again allowed to accumulate, the hyperpnœa becomes intolerable when the carbon dioxide concentration reaches 8–9 per cent., despite the fact that the oxygen concentration is abnormally high. These experiments demonstrate two things : first, that the hyperpnœa of re-breathing is due largely to accumulation of carbon dioxide, and secondly, that a reduction in the concentration of oxygen in the inspired air stimulates breathing.

(3) Another experiment demonstrating the way in which the respiratory centres respond to carbon dioxide and maintain a constant alveolar pCO_2 is to add small quantities of carbon dioxide to the inspired air. The typical effect is shown in Fig. 4. 13. The smallest effective concentration of carbon dioxide in the inspired air is about 1 per cent. (7·6 mm. Hg) and this causes a measurable increase in respiratory minute volume, although the subject is unaware of it. As the carbon dioxide concentration of the *inspired air* is increased the respiratory response also increases until at about 3 per cent. his respiratory minute volume is double the normal value when he is breathing room air. It will be noted in Fig. 4. 13, however, that breathing this concentration of carbon dioxide failed to cause a measurable increase in the alveolar pCO_2. As the inspired carbon dioxide concentration is raised still further,

larger increases in respiratory minute volume occur and a rise in alveolar pCO_2 is also evident. But the point to be emphasised here is that the alveolar pCO_2 breathing 6 per cent. carbon dioxide is very much less than the value would have been, had no increase in pulmonary ventilation occurred. In this connexion, J. S. Haldane found that changing his own inspired gas from room air to 3·8 per cent. carbon dioxide increased his pulmonary ventilation by 258 per cent., but the alveolar carbon dioxide concentration rose only from 5·62 to 5·97 per cent., equivalent to an increase in alveolar pCO_2 from 40 to 42·5 mm. Hg.

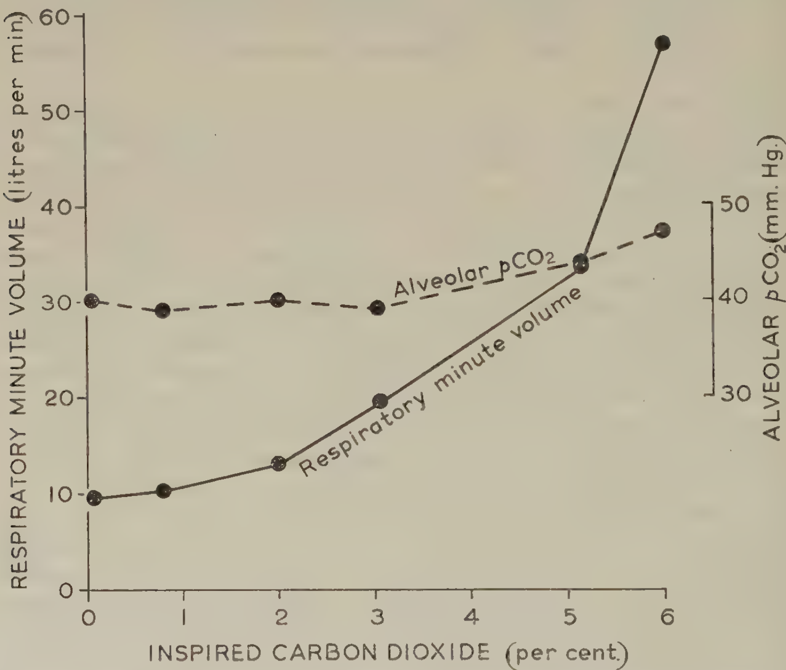


FIG. 4. 13. The effect of increasing the concentration of carbon dioxide in the inspired air on the respiratory minute volume (continuous line) and alveolar pCO_2 (interrupted line) of a man.

Without such an increase in pulmonary ventilation, his alveolar CO_2 concentration would have risen to about 9 per cent. which is equivalent to an alveolar pCO_2 of about 64 mm. Hg.

This experiment demonstrates that the respiratory centre is very sensitive to carbon dioxide, an increase of only 1·6 mm. Hg in alveolar pCO_2 being sufficient to double the respiratory minute volume. It should be emphasised, however, that the normal carbon dioxide content of the atmospheric air (0·04 per cent.) is too small to have any measurable effect on respiration and is in no way responsible for the maintenance of normal respiration.

The importance of the *partial pressure* of carbon dioxide rather than

its percentage can be seen by examination of Fig. 4. 14. Here can be seen the effect of alterations in the barometric pressure, both above and below the normal value, on the alveolar pressure and percentage of carbon dioxide, and on the alveolar pressure of oxygen. It will be observed that while the alveolar oxygen pressure falls steadily as the atmospheric pressure falls, as might be expected, the percentage of carbon dioxide rises steadily and in such a way as to maintain the

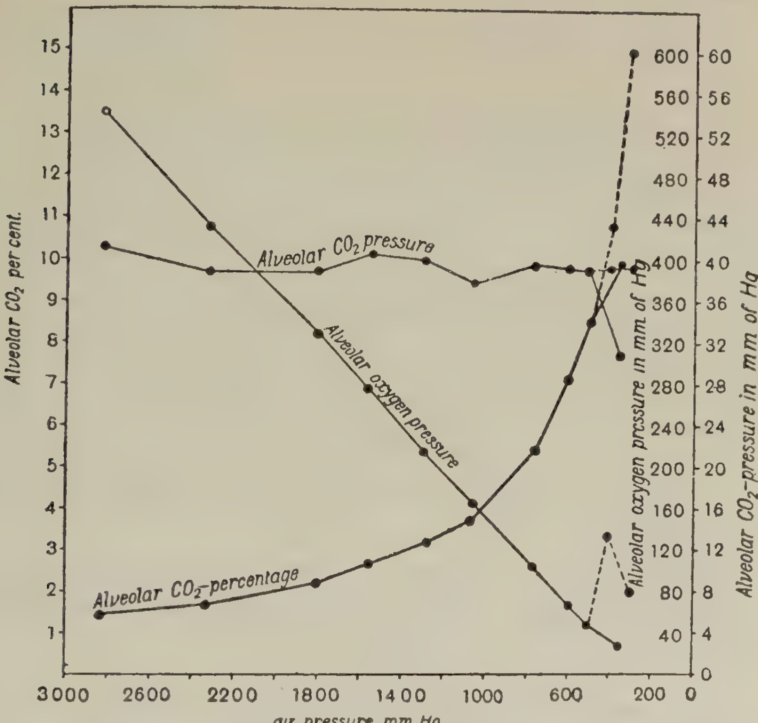


FIG. 4. 14. The effect of alterations in the barometric pressure on the pressure of carbon dioxide, the percentage of carbon dioxide, and the pressure of oxygen in the alveolar air of a man.

The dotted lines show the results obtained when oxygen was added to the air breathed. Note the constancy of the carbon dioxide pressure until the barometric pressure was reduced to values less than 500 mm. Hg (corresponding to a height of about 10,000 ft.), and the respiration was stimulated by oxygen lack. (Boycott and Haldane.)

alveolar carbon dioxide pressure constant. This relation breaks down when the atmospheric pressure falls below 500 mm. Hg (corresponding to an altitude of about 10,000 ft. and an oxygen pressure of 14 per cent. of an atmosphere), owing to the stimulating action of oxygen lack.

The Effect on Breathing of a Reduction in the Partial Pressure of the Carbon Dioxide in the Alveolar Air. A fall in the carbon dioxide partial pressure in the lungs can be brought about by forced breathing. If the

subject is suitable a tracing of his subsequent respiratory movements will follow a pattern similar to that in Fig. 4. 15. In this figure there are three graphs. The lowest is a tracing of the respiratory movements ; the upper and middle curves represent the pressures of the oxygen and the carbon dioxide in the alveolar air. It can be seen from examination of the respiratory tracing that the period of forced breathing was followed by cessation of breathing (apnœa). Then followed a period during which a few breaths were taken, and then a second apnœic

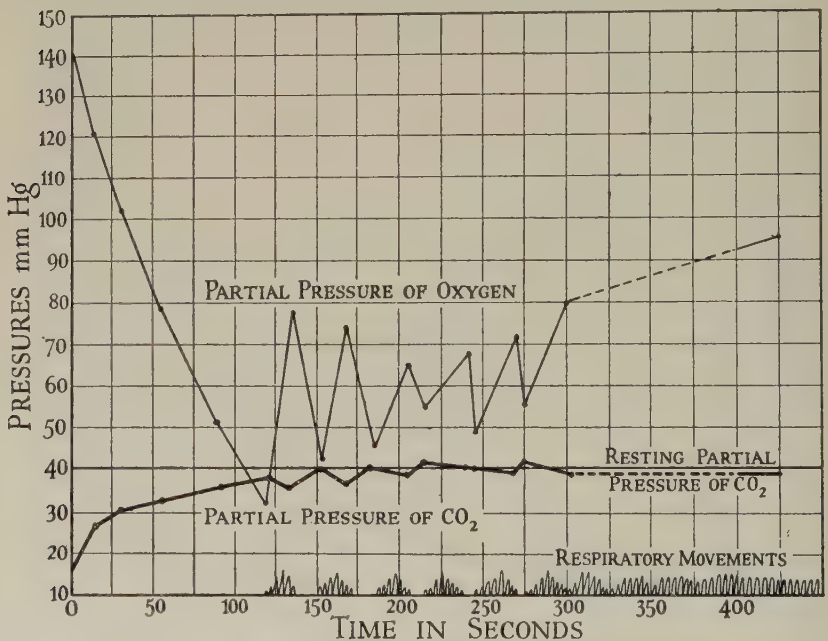


FIG. 4. 15. The effect of **Voluntary Hyperpnœa** on the composition of the alveolar air and on the frequency and depth of respiration.

Voluntary hyperpnœa was carried out for two minutes before the beginning of the record. Note that the first group of respirations begins at a moment when the alveolar carbon dioxide pressure is definitely less than the normal value, owing to the low value of the alveolar oxygen pressure. Similarly, respirations cease before the alveolar oxygen pressure has reached the normal value, owing to the low alveolar carbon dioxide pressure, the result being the production of **periodic (Cheyne-Stokes) respiration.** (Douglas and Haldane.)

period. This " periodic breathing " continued with shortening periods of apnœa until normal breathing was resumed. Amongst the interesting questions which we have to answer are :—

Why are respiratory movements inhibited after forced breathing ?

Why is there a waxing and waning rhythm or periodicity ?

The period of respiratory inhibition or apnœa which followed the period of forced breathing was due to the inhibitory effect of the low alveolar pCO_2, for the following reasons.

(1) During the forced breathing, the carbon dioxide was washed out of the lungs and its pressure was lowered. Examination of the curve in Fig. 4. 15, representing the changes in the alveolar pCO_2, demonstrates that in this subject after the period of forced breathing, the carbon dioxide pressure fell to 15 mm. Hg—well below his normal pressure of 40 mm. Hg—and thus exerted an inhibitory effect on breathing.

(2) If such a subject repeats the forced breathing with a gas mixture containing carbon dioxide (approximately 4·5 per cent.), apnœa will not develop since the alveolar pCO_2 remains at approximately the resting value.

(3) It can be shown that the rise in alveolar oxygen pressure is not to blame, since gas mixtures containing a high percentage of oxygen have no inhibitory effect on breathing.

The explanation of the " periodic breathing " is as follows. During the first period of apnœa, the oxygen pressure fell to a level where it acted as a respiratory stimulus, despite the fact that the carbon dioxide pressure was still below normal, and would thus have exerted an inhibitory influence. It was to satisfy this oxygen want that the subject began to breathe again. A few breaths sufficed to satisfy the oxygen requirements, so that once again the low alveolar carbon dioxide pressure could exert its inhibitory influence, and thus a second apnœic period followed. This periodicity continued until the alveolar pressures of oxygen and carbon dioxide returned to normal.

Reflexes from Chemoreceptors. Chemoreceptors are situated in two locations : in the carotid body at the bifurcation of the common carotid artery and in the aortic bodies in the region of the arch of the aorta. They are supplied by the carotid sinus nerve, a branch of the glossopharyngeal, and by the aortic nerve, a branch of the vagus, respectively. These receptors are not to be confused anatomically with the pressoreceptors (baroreceptors) in the carotid sinus and arch of the aorta (p. 42).

The carotid and aortic bodies consist of epithelioid cells which are surrounded by a rich network of sinusoidal blood vessels (Fig. 4. 16). For its size, the carotid body has a greater supply of blood than any other organ of the body : the blood flow, 2,000 ml./min through each 100 g. tissue, is about four times that of the thyroid gland and 40 times that of the brain (compare Table 2. 3, p. 57).

The chemoreceptors, as their name implies, are sensitive to changes in the chemical composition of the arterial blood, viz., they are stimulated by a decrease in oxygen pressure, and increase in carbon dioxide pressure and by an increase in the hydrogen ion concentration of the blood. They play an important part in the control of breathing under various conditions.

The activity of the chemoreceptors may be studied by recording the electrical activity in fibres of the carotid sinus or aortic nerve. Such studies indicate that, in anæsthetised animals breathing room air, the chemoreceptors are only very slightly active (Fig. 4. 17A). If the arterial

oxygen pressure is normal, then lowering the arterial blood pCO_2, by hyperventilating the animal, from its normal value of 40 mm. Hg to 30–35 mm. Hg, will abolish this normal chemoreceptor activity.

If, however, the animal is made to breathe a gas mixture low in oxygen content, a marked increase in chemoreceptor discharge occurs (Fig. 4. 17B). Of the four types of hypoxia (hypoxic, anæmic, stagnant and histotoxic, Fig. 4. 23, p. 134), anæmic hypoxia is the only one which

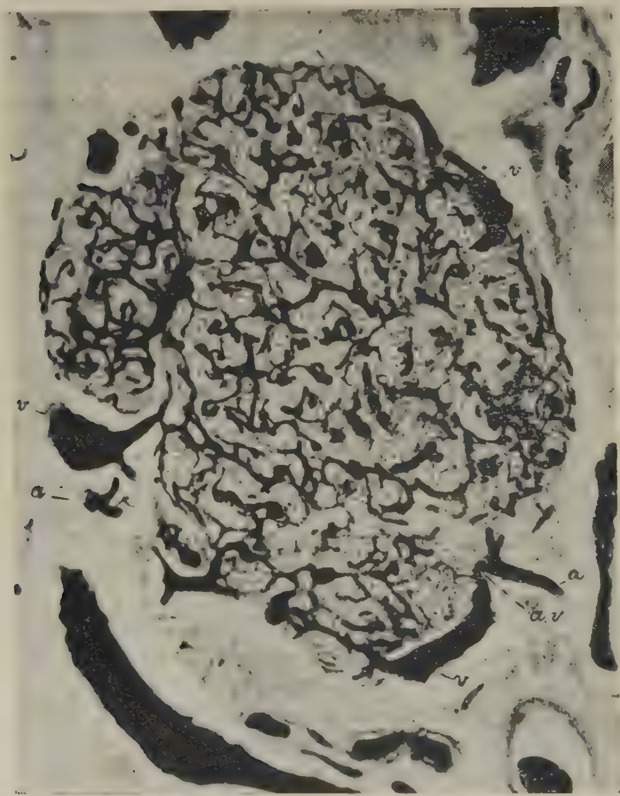


FIG. 4. 16. Carotid body of the adult cat injected with gelatine-carmine showing enormous blood supply, arteries (a), veins (v) and arterio-venous anastomoses ($a.v.$). (De Castro, 1940.)

is incapable of stimulating the chemoreceptors. This is because the receptors are sensitive to changes in the pressure of oxygen in the arterial blood, and not to a lowering of oxygen content alone. Thus, in anæmia produced by lowering the hæmoglobin content or by administration of about 1 per cent. carbon monoxide, which combines with hæmoglobin, there is no chemoreceptor discharge. Stagnant hypoxia, locally in the carotid bodies themselves, on the other hand, produced by lowering the blood pressure and hence carotid body blood flow, and histotoxic hypoxia

A

B

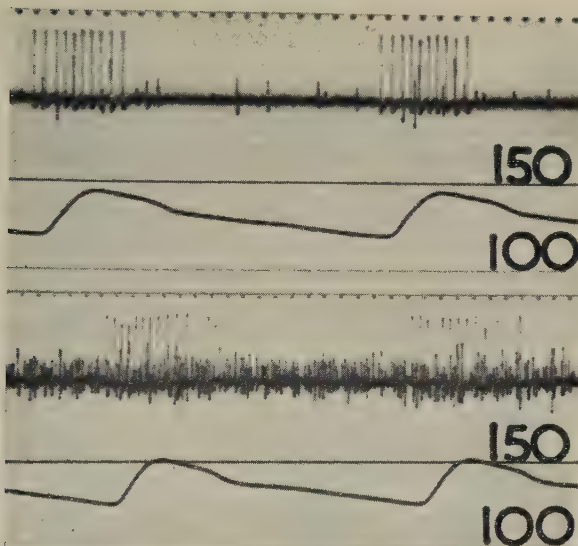


Fig. 4. 17. Afferent impulses in a few fibres of the carotid sinus nerve. The large action potentials are from a single baroreceptor fibre firing synchronously with the anacrotic wave in the blood pressure record; the small potentials are those from chemoreceptor fibres. Cat spontaneously breathing air in A, and 10 per cent. O_2 in N_2 in B. Note the increase in chemoreceptor activity during hypoxia. Blood pressure calibration in mm. Hg. Time trace, 50 cycles per second. (Heymans & Neil, 1958.)

produced by injections of small amounts of cyanide, evoke an intense chemoreceptor discharge.

If the pressure of carbon dioxide in the blood is increased by raising the concentration of carbon dioxide in the inspired air, an augmented chemoreceptor discharge is observed.

Reflexes from the carotid bodies may be studied by isolating the chemoreceptors from the circulation and perfusing them with arterial blood from a donor animal (Fig. 2. 9, p. 48). Changing the gaseous composition of the blood in the donor, and hence in the recipient animal's carotid bodies, evokes changes in respiratory minute volume in the recipient. Lowering the blood pO_2, or increasing the blood pCO_2 or H^+ ion concentration, causes an increase in the rate and depth of breathing. Denervation of the chemoreceptors by cutting the carotid sinus nerves abolishes the response, indicating that it is reflex in nature.

Alternatively, if in an intact animal the arterial blood pO_2 is lowered by administering a gas mixture with reduced oxygen content, an increase in respiratory minute volume results. This hyperpnœa is due entirely to a *reflex* stimulation via the chemoreceptors, for inhalation of the same gas mixture following section of the afferent nerves supplying these receptors causes only depression of respiration (Fig. 4. 18). This latter effect is due to a direct action on the respiratory centre.

On the other hand, inhalation of gas mixtures containing small quantities of carbon dioxide, say 5 per cent., also causes hyperpnœa, but this response is unaffected by section of the carotid sinus nerves, as indicated in Fig. 4. 18. The reflex mechanism, therefore, appears to play only a small part in the control of respiration by changes in arterial blood pCO_2, the main action being a direct one of carbon dioxide on the medullary respiratory centres.

The mechanism of the effect of carbon dioxide on respiration requires a little more explanation in the light of studies of action potentials in chemoreceptor fibres, which have clearly shown that the impulse discharge increases as the arterial blood pCO_2 is raised. How is it, therefore, that interruption of these impulses does not modify the *respiratory*

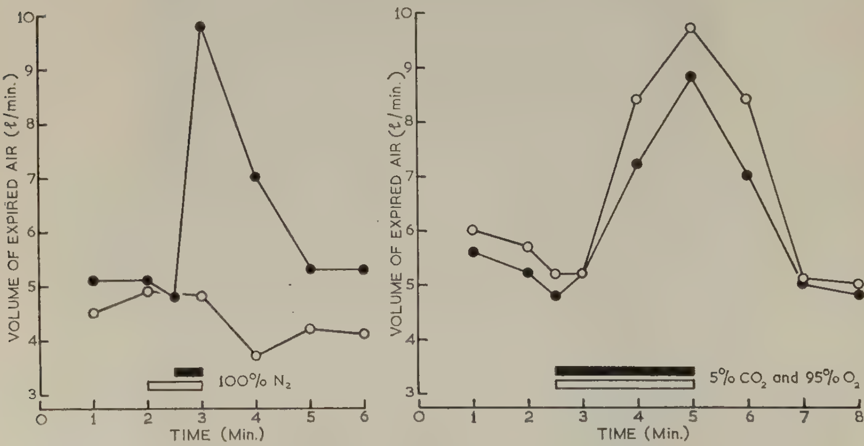


FIG. 4. 18. The effect of denervation of the carotid sinuses on the respiratory response of an unanæsthetised dog to inhalation of nitrogen and of carbon dioxide. Both experiments were made on the same animal: ●, normal, O denervated. (Gemmill and Reeves.)

response to hypercapnia. The position may be summarised by saying : (1) the chemoreceptor discharge produced by raising the arterial blood pCO_2 is relatively small compared to that caused by lowering the arterial blood pO_2 by inhalation of 7–10 per cent. oxygen in nitrogen ; (2) the main effect of carbon dioxide is on the respiratory centres, and this overshadows the reflex component; and (3) increasing the arterial blood pCO_2 of itself tends to diminish the effects of reflexes by an action on central synapses. Thus the small discharge of impulses from chemoreceptors evoked by inhalation of carbon dioxide become blocked, and therefore ineffective, somewhere along the reflex nervous pathway, probably at the centre itself.

Role of the Chemoreceptors in Respiratory Depression. Certain drugs, such as morphine and the barbiturates, depress the respiratory centre and the respiration becomes very much slower. There is an increase in depth, but this does not fully compensate for the decrease in frequency, and the

respiratory minute volume is diminished. In consequence, there is a rise in the alveolar pCO_2 and in the arterial blood pCO_2, and a fall in the alveolar and blood pO_2. The sensitivity of the respiratory centre to carbon dioxide is reduced and it can be shown experimentally that if a subject is given morphine, his respiratory response to an increase in the carbon dioxide concentration in the inspired air becomes smaller than it was before.

In large doses, the sensitivity of the respiratory centre to carbon dioxide may be so depressed that respiration is maintained only by the chemoreceptor reflex mechanism stimulated by the reduction in arterial blood pO_2. In these circumstances, administration of oxygen may have the unexpected, and undesirable, effect of stopping respiration altogether, due to withdrawal of this reflex drive to the respiratory centre.

The Mode of Action of Carbon Dioxide and Oxygen on the Respiratory Centre and Chemoreceptors. The delicacy of the response of the respiratory centre to a change in the carbon dioxide pressure in the alveolar air was demonstrated many years ago by Haldane and his co-workers. Winterstein, later, suggested that the stimulating effect was not due to the presence of carbon dioxide as such, but to the associated rise in hydrogen ion concentration. Mellanby and others, however, found that in animals, injection of lactic acid and sodium bicarbonate caused considerable changes in the acidity of the blood and relatively small changes in breathing, while inhalation of carbon dioxide caused a large increase in pulmonary ventilation and little effect on the hydrogen ion concentration of the blood. In man, also, it has been found that ingestion of ammonium chloride results in a smaller increase in ventilation for a given change in blood acidity than does inhalation of carbon dioxide. In order to account for these and other observations, Gesell pointed out that it was probably the hydrogen ion concentration of the cell interior which was responsible for the control of the activity of the respiratory centre, rather than that of the blood ; it is known that the two may differ appreciably.

Now carbon dioxide is known to penetrate cell membranes very rapidly, whereas ions such as HCO_3^- and H^+ pass through some hundreds or even thousands of times more slowly, and may, indeed, be incapable of penetrating at all. A change in the partial pressure of carbon dioxide in the blood, therefore, would change the conditions within the neurones of the respiratory centre or the carotid body very rapidly ; whereas a change in the concentration of HCO_3^- or H^+ ions would only gradually reach the cell interior. A complete study of the problem, however, cannot be made without quantitative data, much of which is not yet available. It can be stated as reasonably certain at the moment, that the activity of the respiratory centre is determined by a combination of the partial pressure of carbon dioxide and of the concentration of either bicarbonate or hydrogen ions (the two are interrelated), within the neurones or in their close vicinity. The problem is one of great complexity, particularly since it has been shown that acetylcholine possibly plays some part in the excitation of the respiratory centre, as it does in the excitation of so many other neurones.

The mode of action of oxygen is even more speculative. Its depressant action on the respiratory centre might be due to a reduction in the oxidative metabolism, and hence of carbon dioxide production, by the neurones themselves. Alternatively, it is well known that the semipermeable properties of the cell membranes are only maintained in the presence of an adequate oxygen supply ; isolated nerves, for example, lose their excitability in the absence of oxygen, although a much more complete lack of oxygen is needed than that which is sufficient to affect the respiratory centre. In contrast is the stimulating action of oxygen lack on the carotid bodies. This might be due to the production of lactic acid by the neurones, instead of carbon dioxide. The lactate ions would diffuse out very much more slowly than the corresponding amount of carbon dioxide, and the hydrogen ion concentration of the cells would thus be increased.

The Part Played by Afferent Impulses from the Lungs in the Control of Breathing. Vagal afferent nerve endings are present in the bronchi and bronchioles. These nerve endings are sensitive to distension of the lungs, but are insensitive to changes in the partial pressures of oxygen and carbon dioxide.

In 1868 Hering and Breuer showed that interruption of breathing by blocking the respiratory passages during inspiration and expiration had marked effects upon the pattern of breathing, but that when the vagus nerves were cut these effects were completely absent. A few years later, in 1889, Head again demonstrated this reflex (now known as the Hering-Breuer reflex), by observing the effect of inflation of the lungs on the movements of the diaphragm in the rabbit. The rabbit is unique in that its diaphragm is so arranged that it is possible to separate that part which is attached to the ensiform cartilage from the remainder,

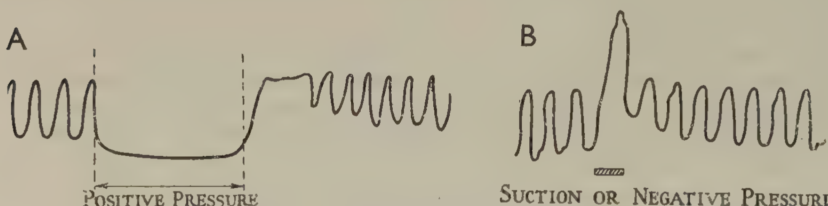


FIG. 4. 19. Recordings of contractions of a slip from the diaphragm of a rabbit. (Upward trend of lever represents contraction of the slip.)

A. Between the arrows a positive pressure was applied to the lungs. The contractions of the slip, and inspiratory movements, were inhibited.

B. A "momentary" diminution in the volume of the lungs (suction) was produced during normal respiration, as indicated. An increased contraction of the slip, *i.e.* an increased inspiratory effort, resulted. (After Head.)

without damaging the blood supply or nervous connections. The active movements of this strip follow exactly those of the remainder of the diaphragm, and can easily be recorded without interference from movements of the chest or diaphragm as a whole, whether active or passive. The essence of Head's results is shown in the accompanying Fig. 4. 19.

In the first tracing (A) a positive inflating pressure has been applied to the trachea, and it can be seen that while the pressure is applied the activity of the diaphragm is inhibited. In the lower figure (B) suction applied for a very short period causes an immediate contraction of the strip of the diaphragm. These effects were abolished when the vagus nerve connections were blocked.

In 1933 Adrian observed in the cat that action potentials were set up in the vagus nerve by inflating the lungs, and that the rate of discharge of the impulses was roughly proportional to the degree of inflation (see Fig. 4. 20). When he recorded the impulses in a single vagal fibre from a cat during normal breathing, he observed that the frequency

waxed and waned in rhythm with the inspiratory and expiratory phases of breathing. The impulses from the lungs fell to zero at the completion of expiration. He also observed that no impulses were recorded when the lungs were deflated (except when such a deflation was far greater than could occur during any normal respiratory movements).

The function of these afferent impulses in the vagus nerve in normal breathing is to signal the depth of inspiration to the respiratory centre and allow expiration to take place, after an adequate tidal volume has ventilated the lungs. Double vagotomy in an otherwise intact

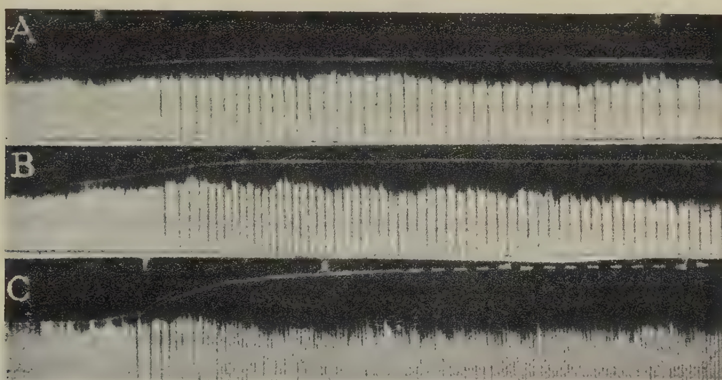


FIG. 4. 20. Oscillograph record of the **Action Potentials in the Vagus Nerve of a Spinal Cat.**

The nerve was cut high up in the neck, and subdivided with fine needles, after removal of the sheath, until only one active fibre remained.

At the top of the record are marks made by a time-signal every quarter second; below this is a white line indicating the position of the lungs, a rise denoting inspiration; at the bottom is the oscillograph record of the action potentials.

The lungs were inflated by a pump, and the frequency of the discharge of the sense organ increases as the inflation increases.

A.	Inflation	65 ml.	Maximum frequency of discharge	80 per sec.
B.	,,	115 ml.	,, ,, ,,	120 ,, ,,
C.	,,	230 ml.	,, ,, ,,	250 ,, ,,

(Adrian.)

animal removes this signal and thus the inspiratory phase is prolonged, and breathing becomes slower and deeper (see Figs. 4. 10 and 4. 12); a normal respiratory minute volume, however, is maintained. Thus when the medullary inspiratory centre transmits its impulses to the inspiratory muscles, it is subjected to an afferent discharge from the lungs which progressively increases in intensity until it inhibits the inspiratory activity and initiates the expiratory phase, allowing the chest wall to collapse in a passive manner. Consequently a cycle of (*a*) inspiratory activity and (*b*) progressive inspiratory inhibition takes place during normal quiet breathing. It has already been pointed out that apneustic breathing occurs, and that rhythmic breathing ceases, when connections between the pontine and medullary centres are cut,

together with section of the vagus nerves. If such a brain-stem section is made when the vagus nerves are intact, rhythmic breathing will continue, due to the part the vagus nerve plays in the cycle of events just described.

To summarise : In *quiet breathing*, the respiratory centre initiates inspiratory activity. Afferent impulses in the vagus nerve signal the degree of expansion of the lungs. This afferent vagal discharge (which increases as inspiration progresses), inhibits the general inspiratory activity and thus the respiratory muscles relax and expiration takes place.

Reflexes from Muscle and Joint Receptors. In experimental animals and in human subjects, passive movements of a limb result in an increase in pulmonary ventilation even when the circulation is cut off by a pressure cuff to prevent metabolic products from the active muscles entering the general circulation and hence the respiratory centre. In man, it was found that passive movement of one leg at the knee 100 times per minute with the circulation occluded increased the respiratory minute volume by 40 per cent. Stimulation of such receptors may play an important part in the increased respiration occurring in muscular exercise.

Protective Reflexes. Stimulation of cutaneous afferent nerves produces an increase in both rate and depth of respiration. Presumably, the afferent fibres which mediate the sensation of pain and temperature are mainly responsible for this increased breathing. Most of us have observed how a cold shower " takes our breath away."

Reflexes arising from stimulation of receptors in the respiratory tract are concerned with the protection of the tract itself. During *swallowing*, respiration is reflexly inhibited by impulses running in the glossopharyngeal nerve from the post-pharyngeal wall. Inhibition of breathing also occurs as a result of stimulation of mucous membranes of the nasal passages by irritant gases. The sensory nerve endings of the trigeminal nerve are involved. In other instances, stimulation of these endings causes *sneezing*, a modified respiratory act. *Coughing* follows similar irritation of the mucous membranes of the pharynx, larynx, trachea and bronchi. The afferent pathway is in the glosso-pharyngeal nerve from the pharynx, and in the vagus nerve from the larynx, trachea and bronchi. The explosive quality of the act of coughing is the result of the initial closure of the vocal cords, which only open after the beginning of the expiratory phase.

The Influence of Changes in Blood Pressure. A rise or fall in arterial blood pressure causes a diminution and an increase in respiratory minute volume respectively. By employing suitable experimental techniques, these changes in respiration have been shown to be due largely to reflexes from baroreceptors in the carotid sinus and arch of the aorta. Such reflexes probably play a minor *rôle* in the nervous control of breathing in the normal animal ; they must not be confused, however, with the very important part played by the baroreceptors in the reflex regulation of the blood pressure (Chapter 2, pp. 42–47).

Adrenaline, in large doses, may in some animals cause a temporary apnœa, or a decrease in the depth of respiration, coincident with the maximum rise in blood pressure. This is in part a reflex response of the baroreceptors of the carotid sinus and arch of the aorta to the rise in pressure, but also due to a cerebral anæmia resulting from an intense vaso-constriction.

The various factors concerned in the nervous control of breathing are summarised diagrammatically in Fig. 4. 21.

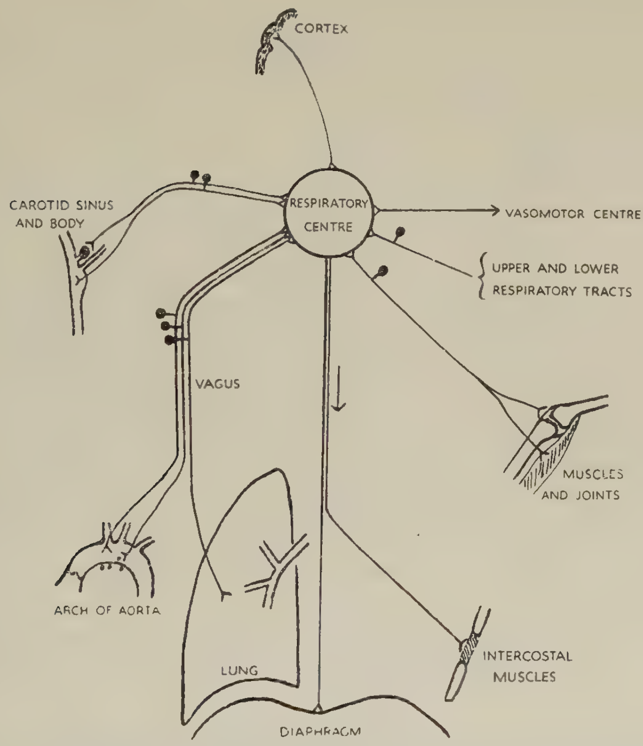


FIG. 4. 21. Diagrammatic representation of the **Nervous Control of Breathing.**

Oxygen Lack

We have seen that when the inspired oxygen concentration is lowered, hyperventilation results, though little change in breathing occurs until the inspired concentration falls to about 14 per cent. (equivalent to 100 mm. Hg) and this corresponds to an alveolar pO_2 of about 60 mm. Hg. When the alveolar pO_2 falls below 60 mm. Hg, the respiratory minute volume increases and this is due entirely to a reflex drive from the carotid and aortic chemoreceptors. However, the chemoreceptors are much more sensitive to a decrease in arterial blood pO_2 than is indicated by the change in respiratory activity disclosed by measurement of the respiratory minute volume. Whereas the respiratory

response to inhalation of gas mixtures of lower oxygen content does not begin until the arterial blood pO_2 falls to 60 mm. Hg (arterial oxygen saturation of 90 per cent.), the electrical activity in chemoreceptor fibres increases with arterial oxygen tensions below 100 mm. Hg (oxygen saturation 96 per cent.). This is due to the fact that the respiratory response evoked by the chemoreceptor drive is antagonised by three opposing factors: (1) a direct depressant effect of hypoxia on the respiratory centre; (2) depression of the respiratory centre due to a lowering

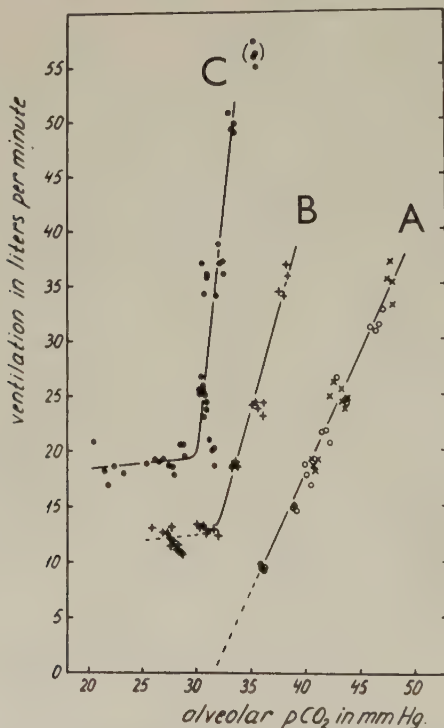


FIG. 4. 22. The relationship between pulmonary ventilation and alveolar pCO_2 (37° C.; prevailing barometric pressure saturated). Alveolar pO_2: (Curve A), X, 168·7 ± 2·1 mm. Hg. O, 110·3 ± 1·9 mm. Hg. (Curve B), +, 47·2 ± 1·5 mm. Hg. (Curve C), ●, 36·9 ± 1·3 mm. Hg. (Nielsen and Smith, 1952.)

of the arterial blood pCO_2 by washing out of carbon dioxide through the reflex hyperpnœa (p. 119); and (3) alkalosis resulting from an increased amount of reduced hæmoglobin. Partially reduced hæmoglobin is able to mop up more hydrogen ions than is oxyhæmoglobin and the effect of the slight alkalotic change in the blood antagonises the chemoreceptor drive. If the hypoxic stimulus continues for some days (as in mountaineers, p. 135), the alkalosis is compensated by the excretion of a more alkaline urine by the kidneys, thereby restoring the normal blood *p*H. Thus, the respiratory response to hypoxia would be much

greater if these three antagonistic mechanisms could be prevented. We can illustrate the effect of maintaining the alveolar pCO_2 constant during hypoxia by taking the following example. A human subject breathing room air has a resting ventilation of 9 l./min and a normal alveolar pCO_2 of 36 mm. Hg (see lower part of curve A in Fig. 4. 22). He is then made hypoxic by lowering his alveolar pO_2 to about 47 mm. Hg and as a result his respiratory minute volume increased to 13 l./min and his alveolar pCO_2 fell to about 27 mm. Hg (see lower part of curve B). His alveolar pCO_2 was then raised by adding carbon dioxide to his inspired air, the alveolar pO_2 being maintained at the same level as before. It was found that his breathing was unaffected until the alveolar pCO_2 was artificially raised to 30–33 mm. Hg, then with further increases in alveolar pCO_2, the respiratory minute volume increased linearly (curve B). When the alveolar pCO_2 had reached its normal level of 36 mm. Hg the respiratory minute volume was 36 l./min. This experiment demonstrates that the ventilatory response to hypoxia is much greater if the concomitant reduction in alveolar pCO_2 is prevented by administering carbon dioxide.

Figure 4. 22 shows one other interesting phenomenon. Curves A, B and C represent the relationship between respiratory minute volume and the alveolar pCO_2 at three different levels of alveolar pO_2, 169 and 110, 47, and 37 mm. Hg, respectively. It will be noted that the slopes of curves B and C, in acute hypoxia, are greater than that of curve A, in normal conditions of air breathing. This means that the sensitivity of the respiratory mechanism to carbon dioxide is greater during acute hypoxia than during air breathing. In other words, the effects of oxygen and carbon dioxide excess are not simply additive, but there is a positive interaction between them, and this is of some physiological importance in asphyxia when hypoxia and hypercapnia occur together.

Asphyxia

The behaviour of an animal when its normal respiratory exchange is seriously interfered with, as by closing the respiratory passages by clamping the trachea or other means (throttling), or by making it rebreathe its own expired air (suffocation), is of some importance. At first, the effect is to produce hypernœa of the normal type, but as the alveolar carbon dioxide pressure rises and the oxygen pressure falls, the respiratory efforts become more and more violent, especially during expiration, and the animal loses consciousness. The second stage is reached when these efforts spread throughout the body ; there is a general excitation of the sympathetic system, the blood pressure rises owing to peripheral vasoconstriction, and practically every muscle in the body takes part in the fight for breath. Soon, however, this gives way to the third stage, that of exhaustion, the medullary centres are *in extremis*, the respiration is reduced to a series of violent inspiratory gasps, the blood pressure falls, mainly due to failure of the asphyxiated heart, the pupils become dilated, and no reflexes can be elicited. Four or five minutes later, the heart stops and the animal is dead.

The Hyperpnœa of Muscular Exercise

The respiratory minute volume increases with the demand for oxygen ; in fact, there is a linear relationship between oxygen consumption and the volume of air breathed when the work is moderate in amount. For any given amount of work, there are two phases in the response of respiration. There is an initial rapid increase in breathing which occurs before any metabolites could reach the respiratory centre and is probably due to irradiation of nerve impulses from higher centres. This is followed by a slow rise in respiration to its final value which depends on the severity of the exercise. Maximum values for the respiratory minute volume are of the order of 80 to 120 litres per minute.

This second phase is difficult to explain. Ever since Haldane stressed the importance of the chemical control of respiration by carbon dioxide, much attention has been given to its possible *rôle* in the production of the hyperpnœa of muscular exercise. The increase in ventilation has been attributed to the rise in pCO_2 and reduction in pH of the arterial blood due to the increased oxidation and to the accumulation of lactic acid. Such a view is untenable, however, because the hyperpnœa sometimes occurs without any demonstrable alteration in the arterial blood pH, pO_2 or pCO_2. Furthermore, the maximum ventilatory response to increasing the inspired carbon dioxide concentration in the resting subject is considerably smaller than that produced by severe exercise.

On the other hand, more recent evidence strongly suggests that in exercise, a lowering of the stimulus threshold to carbon dioxide is an important mechanism responsible for the hyperpnœa. For instance, the respiratory response to inhalation of small quantities of carbon dioxide is much greater immediately after a period of exercise than at rest when this is not preceded by exercise.

Other factors probably contribute to the hyperpnœa. These are :

(1) *Changes in arterial blood pO_2.* It has been shown that in moderately severe exercise the arterial blood pO_2 may fall from 100 to 73 mm. Hg, and this would produce a hypoxic stimulus to the respiratory centre via the chemoreceptors. The pO_2 falls probably because of a combination of two factors : (i) mixed venous blood reaching the pulmonary capillaries has a much lower oxygen content and pressure in exercise than in the resting subject, and (ii) the cardiac output is greatly increased and in consequence the length of time a red blood corpuscle spends in a pulmonary capillary is much reduced.

(2) *Body Temperature.* During exercise, the body temperature rises and this not only stimulates respiration itself, but causes sensitisation of the respiratory centre to arterial blood pCO_2.

(3) *Reflexes.* Several reflexes are known to cause stimulation of breathing : (i) impulses from proprioceptors in the limbs (p. 128) ; (ii) impulses from peripheral chemoreceptors (see section (1) above) ; (iii) impulses from the thoracic viscera ; (iv) impulses from chemically sensitive receptors in exercising muscles. Several known reflexes

arise from receptors situated in the tracheobronchial tree, pulmonary vascular bed, great veins and the heart, and stimulate respiration; and several agents known to be released from exercising muscles, viz. hydrogen, potassium and phosphate ions, cause reflex hyperpnœa when injected into a femoral artery of a decerebrate cat or dog. Little is known, however, about the significance of these two effects in muscular exercise.

(4) *Humoral effects.* Small doses of adrenaline have a stimulating effect on respiration in man, though again, what part, if any, this might play in the production of the hyperpnœa of muscular exercise is uncertain (cf. effect of large doses of adrenaline, p. 129).

Pulmonary ventilation is adjusted closely to the metabolic needs of the body so as to maintain a perfect, or near perfect, homeostasis. In all probability, this is brought about not by any one single factor, but by a collaboration of several contributory causes. In conclusion, it may be said that the most important factor causing hyperpnœa of muscular exercise is the lowering of the threshold of the respiratory centre to carbon dioxide as a result of the acidosis produced by accumulation of lactic acid and the raised body temperature. An additional stimulus in some way connected with want of oxygen appears to be involved as well. We must probably assign only a minor *rôle* to the nervous factors.

Hypoxia

Whenever for any reason the cells of the body do not have, or are unable to utilise, sufficient oxygen to carry on normal function, they are said to be suffering from an oxygen deficiency or hypoxia. Hypoxia may be classified into four types: hypoxic, anæmic, stagnant and histotoxic.

(1) *Hypoxic hypoxia* occurs when there is defective oxygenation of the blood in the lungs causing a reduction in the arterial oxygen pressure. This may result from two causes: (*a*) reduction in the partial pressure of oxygen in the inspired air such as occurs from the addition of an inert gas such as nitrogen or to a fall in the total atmospheric pressure associated with ascents to high altitude. (*b*) Reduction in respiratory minute volume through depression of the respiratory centre by various drugs including anæsthetics, or paralysis of the respiratory muscles.

Under normal conditions, as shown in Fig. 4. 23, the blood leaves the lungs about 97 per cent. saturated with oxygen and reaches the tissues containing about 19·5 volumes per cent. of oxygen (at a pressure of 80 mm. Hg). Here about 5 volumes per cent. are abstracted during rest, so that mixed venous blood contains about 14 volumes per cent. of oxygen (around 70 per cent. saturated and at a pressure of 35 mm. Hg). The gradient between arterial and capillary oxygen pressure under normal conditions is therefore about 45 mm. Hg (80 minus 35). During hypoxia when the arterial oxygen saturation is, say, 75 per cent. (oxygen pressure 40 mm. Hg), there is still about 15 volumes per cent. of oxygen available to the tissues. Since the latter require 5 volumes per cent.,

the remaining 10 volumes in venous blood is held at a relatively low partial pressure of about 25 mm. Hg. This results in an arterial-capillary oxygen pressure gradient of only 15 mm. Hg and a lowering of the oxygen pressure at which the cells have to metabolise.

(2) *Anæmic Hypoxia.* This type is caused by hæmorrhage or anæmia. As will be seen in Fig. 4. 23, the arterial oxygen pressure is normal and the hæmoglobin is 95 per cent. saturated. But because the amount of hæmoglobin per unit volume of blood is considerably reduced, the

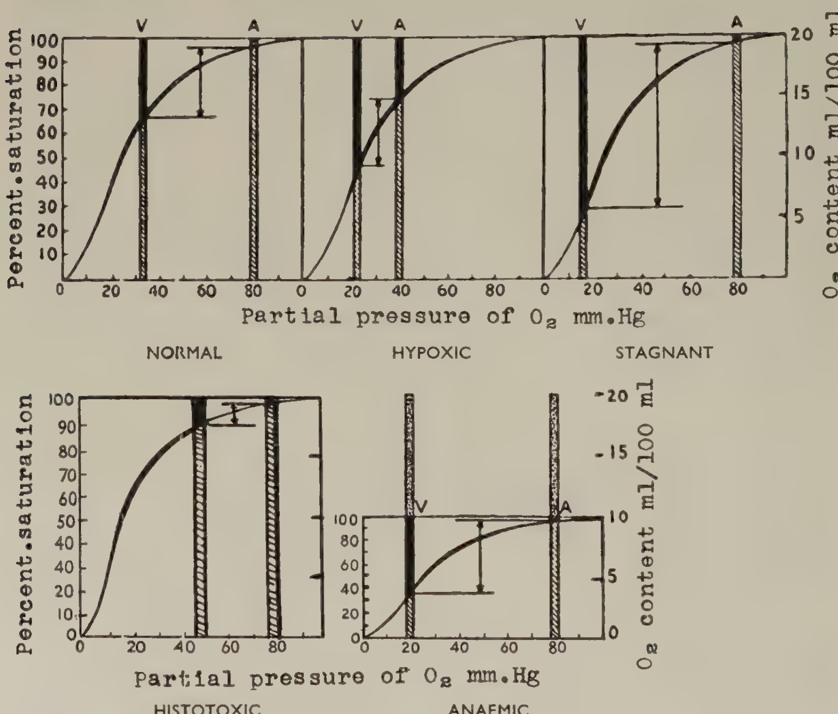


FIG. 4. 23. Diagram illustrating the oxygen dissociation curves in various types of hypoxia. The vertical columns, representing the arterial (A) and venous (V) blood, indicate the amount of oxygenated hæmoglobin (shaded portion) and reduced hæmoglobin (black portion). The perpendicular arrows denote the volume of oxygen per unit volume of blood delivered to the tissues. (See also text.)

oxygen content of the blood is diminished by a similar proportion. A large part of the oxygen supply to the tissues must therefore be delivered at a lower pressure than normal unless the rate of the blood-flow is increased.

(3) *Stagnant Hypoxia* occurs in conditions in which the circulation rate through the tissues is slowed. Although the oxygen content, hæmoglobin saturation and oxygen pressure of the arterial blood are normal (Fig. 4. 23), the venous blood is considerably more reduced than usual resulting in a low tissue oxygen pressure. This is simply due to

the blood spending longer time in the capillaries and, in consequence, a larger volume of oxygen per unit volume of blood is extracted.

(4) *Histotoxic Hypoxia* occurs when the respiratory enzyme systems in the tissues are poisoned, *e.g.* by cyanide. The cells are, therefore, unable to utilise the oxygen carried to them and, as a result, the venous blood contains a high oxygen content (Fig. 4. 23).

The Effects of Low Pressures of Oxygen

At high altitudes, the percentage of oxygen in the atmosphere is the same as at sea-level, but since the barometric pressure is reduced, the partial pressure of oxygen is diminished to a comparable degree. The effects on man of oxygen want depend not only on the altitude to which he ascends, but also on the rate of ascent.

The Mountain Climber. The first signs and symptoms of oxygen lack occur when the healthy mountaineer ascends slowly to about 12,000 ft. above sea-level, corresponding to a pressure of 480 mm. Hg. They are headache, nausea, vomiting and a feeling of lassitude and are regarded as characteristic of " mountain sickness." Mental features such as a feeling of well-being, exhilaration, talkativeness and sometimes emotional outbursts of laughing or crying, and development of fixed ideas may be evident. These effects wear off as the climber becomes " acclimatised " after a few days. The altitude to which man may climb varies with the individual, but without the additional use of oxygen is in the region of 28,000 ft. (close to the summit of Mount Everest). The limit is determined by the pressure of oxygen in his alveoli.

Acclimatisation to Oxygen Lack. (The mountain climber.) When a mountaineer climbs slowly to a high altitude, changes take place both in his respiration and in his blood circulation which are directed towards ensuring an adequate delivery of oxygen to the tissues. This process of adaptation is called acclimatisation. We shall first examine the changes in hæmo-respiratory functions and the process of acclimatisation in a subject who climbs to an altitude of at least 14,000 ft. (barometric pressure, 445 mm. Hg) and remains there.

The Respiration. The first line of defence against oxygen lack at high altitudes is an increase in pulmonary ventilation, and this increase remains even after acclimatisation is complete. The result of this is that the alveolar pO_2 is kept as high as possible. It is those subjects who can produce the greatest increase in pulmonary ventilation, who suffer least from hypoxia and altitude sickness. One of the effects of this hyperpnœa is a fall in the alveolar pCO_2, as is shown in Figs. 4. 14, 4. 15 and 4. 24.

The results of a series of investigations made on persons who were permanently resident at high altitudes, indicate that the alveolar pO_2 falls in proportion to the increase in height, and judging from these results, the pO_2 at 28,000 ft. in an acclimatised subject would be approximately 25 mm. Hg. A sample of alveolar air collected at 22,700 ft.

(310 mm. Hg) had a pCO_2 of 19·3 mm. Hg and a pO_2 of 38·8 mm. Hg. From the values plotted in Fig. 4. 24, the alveolar oxygen pressure of an unacclimatised person at this height would be 35 mm. Hg : acclimatisation thus results in a substantial increase in the pO_2 in the lungs.

Before acclimatisation, breathing is depressed by the low alveolar carbon dioxide pressure, and stimulated by the low oxygen pressure ; breathing is usually " periodic " as described on p. 120. During acclimatisation, there is a rise in the excitability of the respiratory centre, and the kidneys excrete a more alkaline urine, thus reducing the bicarbonate concentration of the blood. The ratio $[H_2CO_3]/[HCO_3]$, and thus the hydrogen ion concentration (Chapter 3, p. 85), returns towards its normal value, and increases the excitation of the respiratory centre. As a result of these changes, the respiratory minute volume becomes greater, in spite of the low alveolar pCO_2, the alveolar pO_2 rises, breathing ceases to be " periodic, " and more oxygen is carried by the blood.

The Circulation. The circulatory system assists in the process of acclimatisation to high altitudes, mainly by an increase in the oxygen-carrying power of the blood. After a stay of four or five weeks' duration at a high altitude, there is an increase both in the number of red blood cells and in the hæmoglobin concentration, due to a response of the blood-forming organs to a reduction in the oxygen pressure of the blood. Permanent residents at 13,000 ft. (4,000 m.) may have red cell counts as high as 8 million/mm^3. The hæmoglobin content of Europeans in the 1952 Cho Oyu expedition averaged 14·9 g/100 ml. at 1,000 ft., and rose 0·8 g/100 ml. per day during ascent to a final average value of 20·3 g/100 ml. The increased amount of hæmoglobin raises the oxygen capacity of the blood and so tends to counteract the lowered arterial oxygen saturation encountered at high altitude.

The increases in heart rate and cardiac output associated with acute hypoxia are well maintained during acclimatisation. The increased blood-flow in tissues is another important adaptive mechanism helping to compensate for the reduction in tissue pO_2 consequent upon the fall in arterial blood pO_2. Redistribution of circulating blood also takes place, so that the more vital tissues will receive a priority of supply.

The above mechanisms are involved in helping to maintain the internal environment of the tissue cells as close as possible to that at sea-level. In this connection it is of interest that at high altitude there is a marked reduction in the gradient between the pO_2 of the ambient air and the pO_2 of the arterial blood. At sea-level, the gradient is $159-100=59$ mm. Hg : at 20,000 ft., for instance, it is $74-40=34$ mm. Hg. There is a corresponding reduction in the gradient between the ambient air pO_2 and the mean capillary blood pO_2.

Acute Oxygen Lack in the Aviator. It has already been pointed out that the climber who ascends slowly and has time in which to become acclimatised has been able to reach heights of 28,000 ft. When an aircraft pilot climbs rapidly, there is no time for these processes to take place and it is, therefore, impossible for him to reach such altitudes

without the use of oxygen. Should he expose himself to the atmosphere at 25,000 ft. for as long as ten minutes he is likely to die.

The effects of oxygen want in the aviator first show themselves at about 5,000 ft., by increased breathing. Above 12,000 ft., mental and physical functions are impaired and over-confidence develops. At 18,000 ft., circulatory changes occur, producing an increase in pulse rate and blood pressure. The senses of touch, pain, vision and hearing are

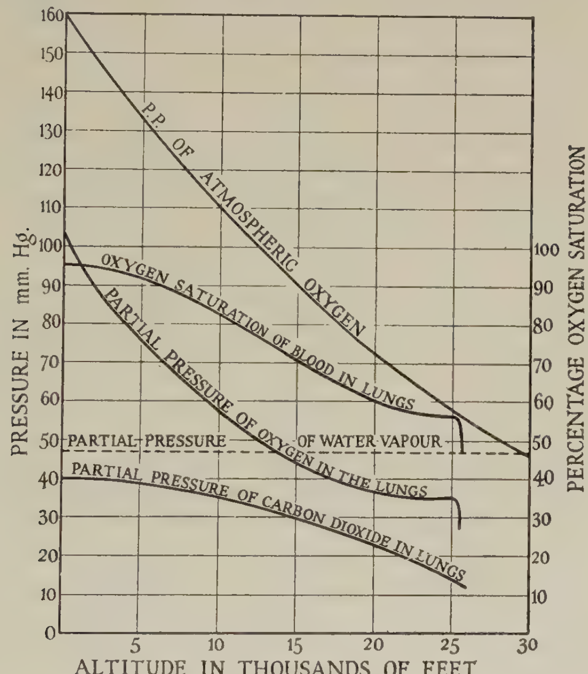


FIG. 4. 24. Curves showing (1) the partial pressure of the atmospheric oxygen, (2) the oxygen saturation of the blood passing through the lungs, (3) the partial pressure of the oxygen in the lungs, and (4) the partial pressure of the carbon dioxide in the lungs ; at all altitudes between sea-level and 30,000 ft. (After Grow and Armstrong.)

impaired. At these altitudes the flyer may observe how much brighter the day appears and how much louder the engines sound when he breathes from his oxygen mask. Between the heights of 18,000 and 30,000 ft., unconsciousness—which is sudden in onset and without warning—followed by paralysis and death may occur.

The effects of oxygen lack are affected by bodily activity. During quiet walking oxygen consumption may be three times that of the resting individual, and the pilot should, therefore, reduce his movements to a minimum.

The changes that occur in the pressures of the lung gases in the high altitude flyer can be seen in Fig. 4. 24. The alveolar air pressures shown were taken at heights up to 25,000 ft., under conditions where

oxygen was not inhaled ; 25,000 ft. is the highest altitude to which a subject can fly without the use of oxygen. The top curve represents the partial pressure of oxygen in the atmosphere ; the next curve represents the oxygen saturation of the blood in the lungs ; and the third curve represents the partial pressure of oxygen in the alveoli. This last curve does not follow a course parallel to that of the atmospheric oxygen pressure because of the hyperventilation caused by the low oxygen pressure ; for the same reason the carbon dioxide pressure (bottom curve) falls, instead of remaining constant. The horizontal broken line represents the water vapour pressure in the lungs which remains constant at 47 mm. Hg.

If the effects of oxygen lack are to be avoided, it is essential that pilots breathe 100 per cent. oxygen above about 10,000 ft. (3,000 m.). This will maintain an adequate alveolar and arterial blood pO_2 up to altitudes of about 40,000 ft. (12,200 m.). At this height the barometric pressure is 140 mm. Hg, the aqueous vapour pressure is 47 mm. Hg, and the alveolar pCO_2, say, 30 mm. Hg, *i.e.* lower than normal due to hyperventilation. The alveolar pO_2, therefore, must be : $(140-47-30) = 63$ mm. Hg. This is an oxygen pressure just sufficient to maintain the pilot orientated and in control of his aircraft, but at altitudes above 40,000 ft., it is necessary to raise the pressure by the use either of a sealed flying suit filled with oxygen under slight pressure, or of a sealed cockpit also under pressure. Using sealed cockpits, oxygen insufficiency ceases to be a limiting factor in flying at high altitudes.

The rarefied atmosphere at high altitudes produces certain other effects upon the body apart from the respiratory effects due to the low oxygen pressure. Speech has a nasal quality, which is partly due to the inability of the rarefied atmosphere to vibrate the vocal chords. Foreign bodies can only be expelled with difficulty, and a cough fails to dislodge particles from the respiratory mucous membranes, so that the cause of the coughing remains and pilots under these circumstances are subjected to continuous irritation. The low barometric pressure causes distension of abdominal organs, which contain gas ; this may cause pain and discomfort. Severe frontal sinus pain may develop for the same reason.

Cyanosis

Cyanosis may be defined as the bluish coloration imparted to the skin by the presence of reduced hæmoglobin in the blood of the superficial capillaries. It depends on the *absolute* amount of reduced hæmoglobin in the capillary blood and not on the relative proportions of reduced hæmoglobin and oxyhæmoglobin. It has been found that 5 g. of reduced hæmoglobin per 100 ml. of capillary blood, or an oxygen unsaturation of 6·7 ml. per 100 ml., is about the threshold level at which cyanosis appears. An anæmic subject, therefore, who has a total of less than 5 g. of hæmoglobin per 100 ml. –*i.e.* an oxygen capacity of less than 6·7 ml. per 100 ml.—does not ordinarily become cyanotic. Thus the presence of cyanosis means that the tissues are hypoxic, although the

absence of cyanosis does not necessarily indicate that there is no hypoxia.

The **causes** of cyanosis are numerous and are most important from the standpoint of diagnosis and treatment of the patient.

(1) **Conditions in which the Alveolar Oxygen Pressure is Reduced.** We have already discussed the problem concerning the mountaineer and the aviator at high altitudes where the lowered oxygen pressure of the air breathed causes hypoxia which may be severe enough to produce cyanosis through a diminished alveolar oxygen pressure (p. 135). Another cause is a reduced respiratory minute volume due either to failure of the respiratory mechanism or to obstruction of the airway. The respiratory centre is depressed, for instance, by overdose of sedative drugs ; peripheral respiratory failure occurs as a result of degeneration of the lower motor neurones to the muscles of respiration as in poliomyelitis. Obstruction may be caused by foreign bodies lodging in the trachea or bronchi. In the case of complete obstruction of a bronchus to one lobe, absorption of air behind the obstruction gradually takes place leading to *collapse* of the lobe. The blood passing through the lobe will not therefore be oxygenated. *Asthma* is another form of partial obstruction of the airways so that the movement of air in and out of the lungs is hindered by contraction of the bronchial muscle.

(2) **Conditions in which there is Impaired Diffusion across the Alveolar Membrane.** This occurs in pathological states in which the alveolar membrane is abnormally thick due to inflammatory processes. The presence of œdema fluid in the alveoli also hinders diffusion and, in consequence, increases the gradient of the oxygen pressure between alveolar air and arterial blood. A similar state of affairs occurs in *emphysema* in which there is destruction of many of the septa between alveoli causing a reduction in the area available for gas exchange.

(3) **Conditions in which there is Abnormal Reduction of Hæmoglobin in the Systemic Capillaries.** This may be due either to an increase in oxygen utilisation by the tissue or to a decrease in blood-flow through it. The latter is the more usual cause of cyanosis in this group and gives rise to stagnant hypoxia (p. 134). It may occur *locally* as a result of intense vasoconstriction through cooling a part of the body, or it may be *generalised* as in advanced heart failure and other conditions in which the peripheral circulation is abnormally sluggish.

(4) **Conditions in which there is an Abnormal Mixture of Arterial and Venous Blood.** In many forms of congenital heart disease, there is cyanosis due in part to direct mixture of venous with arterial blood through abnormal communications between the left and right sides of the heart. In other conditions which have already been referred to, such as obstruction to the air passages involving a part or whole of one lung, arterial blood will be a mixture of oxygenated and venous blood if the pulmonary blood-flow through these areas is maintained. In point of fact, the pulmonary circulation gradually closes down in the collapsed part of the lung thereby reducing the volume of venous blood mixing with arterial blood and hence diminishing the cyanosis.

(5) **Conditions in which there is Alteration in the Hæmoglobin.** A number of drugs and poisons are capable of converting the iron in the hæmoglobin from the di- to the trivalent form causing the formation of *methæmoglobin.* This is dark in colour and so gives rise to cyanosis. In contrast with all the conditions giving rise to cyanosis enumerated in (1) to (4) above, the arterial oxygen pressure in methæmoglobinæmia is normal and, in consequence, there will be no stimulation of respiration reflexly through the chemoreceptors. The oxygen-carrying power of the blood is, however, reduced in proportion to the amount of methæmoglobin present. *Carbon monoxide poisoning* may also be mentioned here, since there is a similar loss of oxygen-carrying power of the blood. There is, however, no cyanosis, since carboxyhæmoglobin is red in colour. Carbon monoxide combines with hæmoglobin with an affinity about 250 times that of oxygen : 0·1 per cent. in the air produces a concentration of 15 to 20 ml. per 100 ml. in the blood. Very small concentrations in the air are therefore sufficient to produce severe symptoms in thirty to sixty minutes (Chapter 3, p. 81).

The Therapeutic Value of Oxygen

In treating certain forms of hypoxia, the administration of high concentrations of oxygen is a measure of the utmost value. By increasing the alveolar oxygen pressure the hæmoglobin saturation of the blood leaving the pulmonary capillaries will be increased in conditions such as those in which the alveolar oxygen pressure is reduced and in which there is impaired diffusion across the alveolar membrane. There is relief of the cyanosis and a lessening of the dyspnœa associated with these conditions. Oxygen is also of extreme value in carbon monoxide poisoning for the high oxygen pressure displaces carbon monoxide from the blood.

In other forms of hypoxia, oxygen therapy is of less value. In anæmic hypoxia, for instance, the blood leaves the lungs with its hæmoglobin fully saturated with oxygen, so the only benefit would be derived from the increased amount of oxygen carried in physical solution in the plasma. For a similar reason, oxygen is usually of less use in stagnant hypoxia.

Oxygen has certain toxic effects, and these are described in the next section.

Respiration at High Atmospheric Pressures

When a man exposes himself to air at high pressures, he may develop certain symptoms either during exposure to the increased pressure or when he has returned to normal atmospheric pressure.

Increased air pressure is met in deep sea diving, in caissons and in submarine escape apparatus (the pressure within the submarine is at one atmosphere). Every 33 ft. of sea-water means an additional pressure of one atmosphere, so that as a diver exposes himself to increased pressure the volume of respiratory gases *dissolved* in his blood plasma and tissues increases in proportion to the raised partial pressures of these gases in

the alveolar air (Henry's law). Thus at two atmospheres pressure, twice as much gas will be dissolved in his blood and tissues as at one. The gases we have to consider are oxygen and nitrogen ; the alveolar pCO_2 remains almost constant, and so carbon dioxide is not a gas which presents serious problems in this connection.

Nitrogen. The dangers of nitrogen under increased pressure arise from two facts : (1) Nitrogen diffuses relatively slowly through living membranes, and (2) it is about five times as soluble in fat as in water. Consequently, not only does it take a considerable time for the body to absorb the extra nitrogen at any given high atmospheric pressure, but the elimination of the extra nitrogen when decompression occurs is also prolonged. If a diver ascends too quickly to a lower pressure, the nitrogen may come out of solution and form bubbles of gas in his blood and tissues, in particular in the central nervous system on account of its high fat content. These bubbles are apt to lodge in capillaries and obstruct the flow of blood giving rise to localised symptoms due to asphyxia and to distension of the tissues. These symptoms are pains in the muscles and joints, loss of cutaneous sensation and paralysis through involvement of the central nervous system (" caisson disease," " bends," " divers' palsy "). The severity of the symptoms depends on the pressure attained, the length of time spent by the subject at that pressure and on the rate of ascent. In severe cases, death may result from multiple air emboli in the heart and brain.

In order to prevent sudden evolution of gas, divers must ascend slowly, so that the tissues have time to get rid of their excess nitrogen via the lungs, without the formation of bubbles. Treatment of a case of caisson disease involves recompression in a pressure chamber, to dissolve the gas bubbles, followed by slow decompression to atmospheric pressure.

Another method of preventing the formation of nitrogen bubbles is to replace atmospheric nitrogen by helium, since helium is an inert gas and less soluble in fat than is nitrogen. In practice the diver is given a helium-oxygen gas mixture to breathe. This has another advantage, namely, it lessens " nitrogen narcosis," a condition associated with the onset of euphoria, hilarity, impaired mental activity and an increased difficulty in concentrating on and performing an allotted task. These effects, though they are very slight, are observed when air is breathed at 3 atmospheres pressure ; they begin to handicap the subject at 4 atmospheres, and may rapidly make a man helpless at 10 atmospheres.

A similar condition of bends may occur in occupants of high-flying aircraft. Machines which are intended to fly at altitudes greater than 40,000 ft. (barometric pressure less than 140 mm. Hg) are fitted with pressurised cabins in which the pressure is maintained at an equivalent altitude of about 8,000 ft. (barometric pressure 560 mm. Hg). This enables an adequate oxygen pressure to be maintained in the lungs of the aviators. Should an accident occur, resulting in the pressure in the cabin being suddenly reduced to the ambient pressure, oxygen and nitrogen bubbles may form in the blood, and cause caisson disease.

Oxygen. Gas mixtures containing up to 60 per cent. of oxygen may be inhaled without danger. Pure oxygen (100 per cent.) breathed for more than twenty-four hours may produce mental dullness and evidence of pulmonary congestion. Newborn infants are particularly susceptible to oxygen poisoning : there is proliferation of the retinal vessels into the vitreous humour with excess formation of fibrous tissue (retrolental fibroplasia), which may lead to permanent blindness. When oxygen is to be used in infants for resuscitation purposes, the concentration should be limited to 40 per cent. in the inspired air.

Inhalation of pure oxygen at four atmospheres pressure produces signs and symptoms of oxygen poisoning which include convulsions and a fall of blood pressure. These were thought to be due to the increased amount of oxygen dissolved in the blood, so that when the blood passes through the tissues, less oxygen is lost from combination with hæmoglobin. As we have seen in Chapter 3, hæmoglobin becomes more acid when it combines with oxygen and is, therefore, less ready to surrender base for carbon dioxide combination. This results in a reduction in the amount of carbon dioxide removed from the tissues so that the effects of oxygen excess may be due to accumulation of carbon dioxide in the tissues. If this is the correct explanation, the carbon dioxide pressure should rise in the venous blood. More recent work indicates that, at these high oxygen pressures, the carbon dioxide pressure in cerebral venous blood, and hence in brain tissue, does rise, but that the rise is small and can in fact be excluded as an important contributing cause of oxygen poisoning in man. A direct toxic effect of oxygen on tissue enzyme systems, in particular those containing sulphydryl groups, is the most likely explanation.

Artificial Respiration

There are many circumstances in which respiratory movements cease temporarily. Death will follow from asphyxia unless fresh oxygen can be supplied, and the excess carbon dioxide washed out, by some means of artificial ventilation of the lungs. The cause of the respiratory failure can often be removed in this way, and complete recovery results. Common instances of such circumstances are drowning, electric shock, asphyxia from smoke, etc., in fires, carbon monoxide poisoning, certain diseases of the central nervous system which result in paralysis of the respiratory muscles, and the failure to breathe of the new-born.

In patients undergoing certain surgical operations it is desirable to produce complete muscular " relaxation " which is done by administering a drug which blocks transmission at the neuro-muscular junction. Since all skeletal muscles, including the diaphragm and intercostal muscles, are paralysed by such drugs, respiration must be maintained artificially.

In the emergencies mentioned above, it is usual for respiration to fail before the heart stops beating. As asphyxia gradually develops, however, not only does it adversely affect the heart, but it also depresses the medullary vasomotor centre. It is absolutely essential, therefore,

that oxygen should be supplied to the tissues as quickly as possible to prevent deterioration of the heart and circulation, and to give the respiratory centre every opportunity of recovering its normal rhythmic activity. Before starting artificial respiration in such emergencies, however, it is necessary first to ensure that the patient's airway is not obstructed ; no time should be wasted before loosening clothing and removing foreign bodies and water from the mouth and upper respiratory tract.

Artificial respiration may be performed in one of the following ways. (1) Air is blown into the lungs (intermittent positive pressure breathing). This may be carried out by means of a respiratory pump connected to a cuffed tube inserted through the patient's mouth into his trachea. Alternatively, a tracheotomy is performed, that is, a cannula is inserted into his trachea through a mid-line incision in the neck. This latter procedure is the method of choice for artificially ventilating anæsthetised animals. (2) The thoracic cavity is distended by applying suction to the outside of the body. In this method the patient is placed with his body in an airtight chamber and his head outside, his neck passing through an airtight collar in the wall. When air is sucked out the chamber by a rhythmically operated pump, the patient's chest expands and air is drawn into his lungs ; the reverse happens when air enters the chamber again during the second half of the " respiratory " cycle. This apparatus is known as the " Iron Lung." (3) The thoracic cage is compressed, either directly by pressure on the ribs, or indirectly by pressure on the abdominal contents and so on the diaphragm. Respiration may sometimes be induced, also, as a result of the nervous reflex set up by pulling out the tongue or applying some other painful stimulus.

Emergency methods. In man, the method of manual artificial respiration now recommended by the League of National Red Cross Societies is known as the arm-lift back-pressure (ALBP) method of *Holger Nielsen* (Fig. 4. 25). The subject lies in the prone position with his arms above his head and elbows flexed so that one hand rests on the other ; his head, turned to one side, lies on the uppermost hand. The operator kneels on his left knee at the subject's head, with his right foot near the subject's left elbow (A). Grasping the subject's arms just above the elbows, the operator rocks backwards, raising the subject's arms until a resistance is felt (B). The arms are then dropped, and the operator, placing his hands just below the scapulæ (C), rocks forwards keeping his arms straight, until his arms are in a vertical position, at the same time maintaining a steady pressure on the subject's chest (D). The movements of lifting and compression occupy about 2½ seconds each and should be repeated about 12 times a minute.

A form of artificial respiration which is becoming increasingly popular is known as the "*Mouth-to-Mouth*" or *expired air* method. In this method, first practised by Elisha (2 Kings, Chap. 4, verse 34), the operator blows air into the subject's lungs intermittently. The operator kneels to one side of the head of the subject who is lying in the supine position. With one hand resting on the subject's forehead and occluding his nose,

the subject's head is fully extended ; the other hand pulls the chin downwards so as to open the subject's mouth. The operator then takes a deep breath and applies his *wide open* mouth to the subject's, and blows air into the subject's lungs. During expiration, the operator withdraws his head so as to enable him to take in another breath and to allow expulsion of air from the subject's lungs by means of the elastic recoil of his lungs and chest wall. The cycle is then repeated about 12 times per minute. The main advantage of this method is that one knows immediately if there is any obstruction in the subject's airway ; furthermore, one can tell from movements of the subject's chest wall approximately how much air is entering his lungs.

When applying artificial respiration, the administration of oxygen

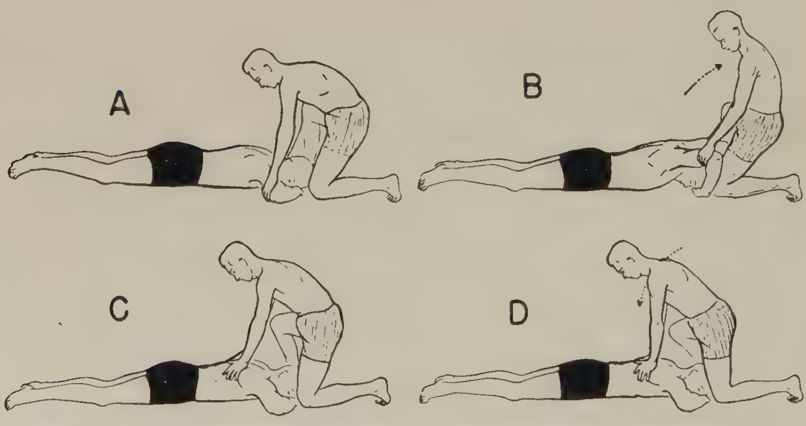


Fig. 4. 25. Holger-Neilsen method of manual artificial respiration. A, placing hands for arm lift. B, arm lift. C, placing hands for back pressure. D, back pressure.

may be beneficial. The addition of small quantities of carbon dioxide, however, is not recommended ; patients requiring resuscitation almost certainly already have a high arterial blood pCO_2, and to administer to them more carbon dioxide may depress the respiratory centres, rather than stimulate them, due to its narcotic action in high doses.

An important consideration in the choice of method of artificial respiration is its effect on the patient's circulatory system. Any method adopting the principle of intermittent positive pressure breathing, whether applied by means of a pump, or by the " mouth-to-mouth " emergency method, causes the mean intrapleural pressure to rise *above* atmospheric pressure and, in consequence, abolishes the " thoracic pump " aiding venous return to the heart. As a result, the cardiac output falls considerably, particularly if the patient is in a state of shock. On the other hand, any method adopting the principle of inflating the lungs by expansion of the thorax *alone* from without, thereby making the intrathoracic pressure more " negative," as for instance in raising

the arms in the Holger Nielsen method, improves venous return to the heart, and hence increases cardiac output. Thus the beneficial effect of applying this method of artificial respiration may lie as much in improving the state of the circulation, and so the blood supply to the brain, as in raising the oxygen saturation of the arterial blood.

CHAPTER 5

DIGESTION

Most of the food we eat is unsuitable for direct use by the cells of the body, either because it is solid or colloidal and therefore cannot pass through the wall of the intestine into the blood, or because, although diffusible, it is in some form which the cells cannot at once assimilate.

The effect of digestion is to resolve the different foodstuffs into simple components which will pass easily through the intestinal mucosa into the circulation and from there, into the cells which are to make use of them. Polysaccharides must be broken down into monosaccharides, fats into fatty acids and glycerol, and proteins into their constituent amino-acids. A few substances, such as the fat-soluble vitamins which have no value as a source of energy but are of vital importance to the body, may have to be rendered soluble in water before they can be absorbed.

This necessary and radical transformation of the food is effected by the enzymes contained in the digestive juices which are poured into the gut by the various glands situated in or near it, whenever food is eaten. It is noteworthy that all these enzymes are *hydrolytic* in their action, and the minimum of energy is wasted during the process of digestion. Absorption of the products of digestion, which proceeds coincidently with their liberation throughout the small intestine, is also a highly efficient process ; the material which is finally collected in the colon for excretion has little food value and consists largely of cellulose, bacteria and débris from the intestinal mucosa.

Besides the products of digestion, the water, salts and organic constituents of the juices themselves must be absorbed. The total volume of digestive juices secreted daily is not accurately known, but it is estimated to be some 4 to 9 litres in man, *i.e.* of the same order as the volume of circulating blood. There is thus a very large daily " turnover " of water and salts between the blood and intestinal lumen, and if anything prevents re-absorption of the fluid (*e.g.* loss through vomiting or diarrhœa) *dehydration* of the body tissues quickly ensues, fluid being withdrawn from these into the blood to maintain its volume.

The Secretory Work of the Digestive Glands. Vasodilatation occurs in all the digestive glands when they are active, as is indicated in Fig. 5. 1, and again in Fig. 5. 3 below, so that during the digestion of a meal there is a great increase in the blood-flow through the portal circulation. A ready supply of water and salts is thus assured for the production of secretions ; but the enzymes and other organic constituents, *e.g.* mucin, are probably prepared from " raw materials " in the blood by the gland-cells themselves. Many of the cells in the digestive glands contain granules or droplets which are apparently

antecedents of organic constituents of the juice. These accumulate during inactivity and are discharged during secretion, particularly if this is prolonged ; these cellular changes can be correlated to some extent with the amounts of enzyme, mucin or other organic material found in the juice (Fig. 5. 2).

The act of secretion is not merely a washing-out of preformed constituents from the cell by fluid filtered off from the blood ; the submaxillary gland can produce saliva against a pressure much higher than that in the arteries (Ludwig) and during secretion its usage of oxygen and sugar is increased (Barcroft).

Apart from the synthesis of organic materials by gland-cells *osmotic*

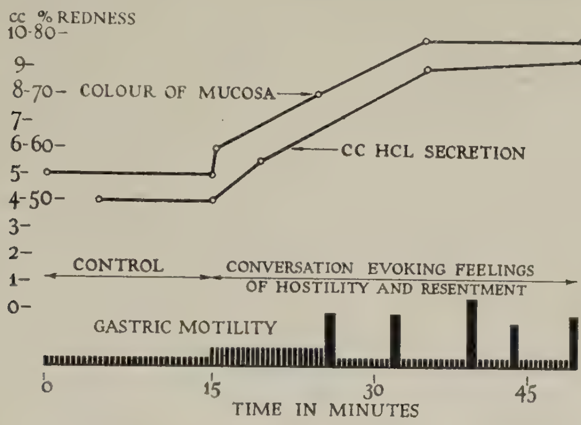


FIG. 5. 1. Contractions of the human stomach (gastric motility) and changes in acid secretion and vascularity (colour) of the mucosa accompanying feelings of hostility and resentment aroused by conversation. The subject was a man (" Tom ") who had a large gastric fistula. The intensity of the red colour of the mucosa was expressed in arbitrary units. The strength of the contractions of the stomach is indicated by the height of the black bars. (Wolf and Wolff, " Human Gastric Function.")

work may be done during secretion ; for instance, the parietal cells of the gastric glands concentrate hydrogen ions about three million times in preparing the acid of the gastric juice from blood.

Innervation of the Digestive Glands and Alimentary Tract. The existence of " secretory " nerves was discovered by Carl Ludwig (1851) who stimulated the lingual nerve, a branch of which (the *chorda tympani*) supplies the submaxillary gland, and found that it caused the secretion of saliva. It has since become abundantly clear that all the digestive glands receive a dual innervation from the autonomic nervous system (Chapter 15), namely, vasodilator and " secretory " fibres from the parasympathetic division, and vasoconstrictor and (possibly) inhibitory fibres from the sympathetic division. The former are distributed to the abdomen in the vagus and pelvic visceral nerves ; the latter for the most part run from the autonomic ganglia to the viscera along the walls of the large arteries.

There is a similar dual nerve supply to the smooth muscle of the alimentary tract, the parasympathetic fibres increasing the motor activity, and the sympathetic fibres depressing it, so that in general the state of activity

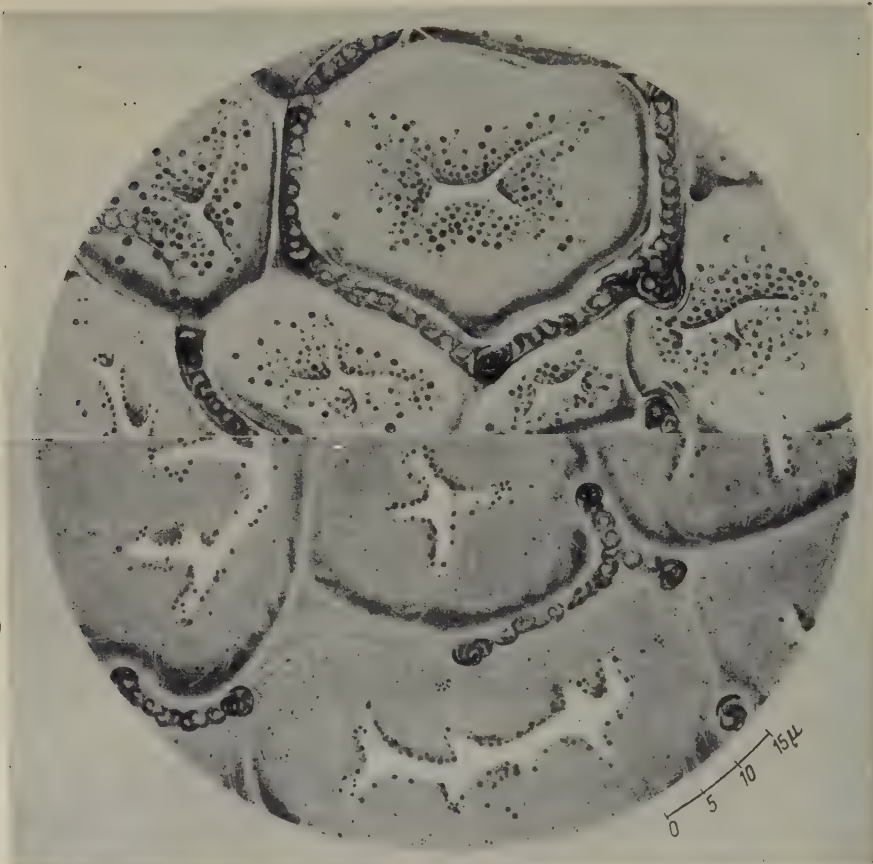


Fig. 5. 2. Portions of the unstained living pancreas of a white mouse (*a*) after twenty-four hours' fast ; the cells contain plenty of zymogen granules, and (*b*) after three hours' stimulation of secretion. Most of the intracellular material has been discharged. (Hirsch.)

of an organ may be said to represent the resultant of the influence of the two systems. However, the extent to which the motor and secretory functions of the digestive tract are normally controlled by these nerves remains problematical ; for instance, complete denervation of an intestinal loop produces a striking increase in tone, motility and spontaneous secretion, but a rapid recovery occurs and in a few days the behaviour of this denervated loop is almost indistinguishable from normal.

In addition to the motor or efferent autonomic fibres mentioned, *afferent* fibres carry sensory impulses from all parts of the tract to the central nervous system ; the reflex arcs formed by these with the autonomic nerves play an important part in the activities of the gut. Sensations from the normally functioning digestive tract are almost entirely absent, apart from fullness of the stomach after a meal and of the rectum before defæcation, which are readily appreciated. Excessive distension or strong contractions of the intestines, particularly in the presence of inflammation or poor blood-supply, gives rise to *pain* which is griping or colicky in character, is poorly localised

and may be " referred " to areas supplied by somatic nerves entering the spinal cord in the same segment as the visceral afferents (see p. 390).

The ultimate " centre " in the brain for visceral afferents and efferents appears to be the *hypothalamus* (p. 460) which thus exerts a general influence over the motor, secretory and vascular reactions of the entire alimentary tract ; damage to or experimental interference with this region of the brain produces changes in secretion and motility of the stomach and intestines.

The hypothalamus has connections with the thalamus and cerebral cortex ; and stimulation or destruction of certain areas of the cerebral cortex causes changes in motility and secretion of the alimentary tract. Furthermore, the thalamus is well known to be concerned with the perception of the painful or pleasurable quality of sensations. Means thus exist by which disagreeable or pleasant emotions may influence the working of the digestive tract. A good example of this is afforded by the experiment depicted in Fig. 5. 1, which is taken from the study by Wolf and Wolff of the daily variations in gastric function of a laboratory technician (" Tom ") ; this man had had from childhood a large gastric fistula, permitting inspection of the interior, withdrawal of contents, etc. It is common everyday experience that pain, fear, anger, resentment or worry are potent causes of " indigestion," and similar upsets of the gastro-intestinal tract.

Saliva

The prompt response of the salivary glands to the sight, smell, or even the anticipation of appetising food, is familiar to everyone as " watering of the mouth." This is a " conditioned " reflex (p. 447), *i.e.* one which has become established by training and experience and in which the cerebral centres play an important part. The stimuli received by the special sense organs are conveyed to the cerebral cortex ; from there they are relayed to cells in the medulla which form the " salivatory nuclei " (they lie in the reticular formation in the floor of the fourth ventricle, see p. 393), and from these cells fibres run to the various glands. Salivation is also brought about when food is actually eaten, by direct stimulation of sensory end-organs in the mucosa of the mouth, tongue and pharynx, from which impulses are transmitted to the salivatory nuclei. This reflex is " unconditioned " ; it is present from birth and does not involve the higher cerebral centres. Thus, reflex salivation is readily produced in a decerebrate cat (p. 408) by introducing acid, alcohol, etc., into the mouth.

Saliva has a *p*H of about 6·8, is fairly well buffered and contains a lubricant *mucin* and (in man) the enzyme *ptyalin* which breaks down starch into a mixture of dextrins and maltose. The main functions of saliva are to moisten and lubricate the food, thus preparing it for swallowing, and to dissolve its soluble constituents so that the flavour is appreciated and the secretion of saliva itself and of other digestive glands thereby stimulated.

Swallowing

Each mouthful of food is chewed and mixed with saliva until it forms a pulpy mass or " bolus " suitable for swallowing, and is collected from time to time on the surface of the tongue for this purpose. Swallowing begins with a quick contraction of the tongue muscles which

propels the bolus past the faucial pillars into the pharynx ; from then onwards, its progress is beyond voluntary control and is accomplished by a rapid and complicated series of movements which constitute the "swallowing reflex" and are co-ordinated by a "centre" in the medulla.

As the bolus enters the pharynx, the soft palate is approximated to the posterior pharyngeal wall by contraction of its muscles, so as to prevent entry of food into the nasal passages. At the same time the larynx is brought upwards and forwards under the shelter of the base of the tongue, raising and opening the upper end of the relaxed œsophagus, which the bolus now enters. The epiglottis may turn backwards to guard the entrance of the larynx ; and the risk of food entering the air-passages is lessened by a reflex approximation of the vocal cords and momentary inhibition of respiration, which also form part of the reflex.

The passage of the bolus from the mouth to the upper part of the œsophagus is so rapid, and the reflex movements so easily disturbed by experimental procedures, that it has proved very difficult so far to elucidate many of the details of the act, and in particular the nature of the forces which transport the bolus so rapidly. X-ray cinematography in subjects swallowing radio-opaque material is proving a most valuable technique for studying the mechanism of swallowing.

The swallowing reflex is touched off by contact of the food with areas on the fauces, pharynx and tonsils which are very sensitive to tactile stimulation and from which impulses travel to the medullary centre ; if these areas are anæsthetised by painting them with cocaine, swallowing is impossible.

Having travelled through the upper third of the œsophagus in a fraction of a second, the bolus is now carried the rest of the way much more slowly by an advancing ring-like contraction of the smooth muscle of the œsophagus. If the bolus is soft and well-lubricated, it reaches the cardiac sphincter at the entrance to the stomach in a few seconds ; but if dry, it may take a minute or so and *secondary waves* (which give rise to a painful sensation in the chest) may arise in the œsophagus and force it along.

Liquids, owing to the impetus given them by the act of swallowing and the effect of gravity, outstrip the œsophageal wave and arrive at the cardiac sphincter in a second or two, where they wait for the arrival of the œsophageal wave. When this approaches the cardiac sphincter the latter relaxes before it and the food enters the stomach.

Gastric Digestion

X-ray examination of the human stomach after eating meals made radio-opaque by the addition of barium sulphate shows that there are wide variations among apparently normal persons in the position, shape and motility of the stomach. An ordinary meal begins to leave the stomach less than thirty minutes after it is eaten, and although the rate of gastric emptying varies with the size of the meal and its

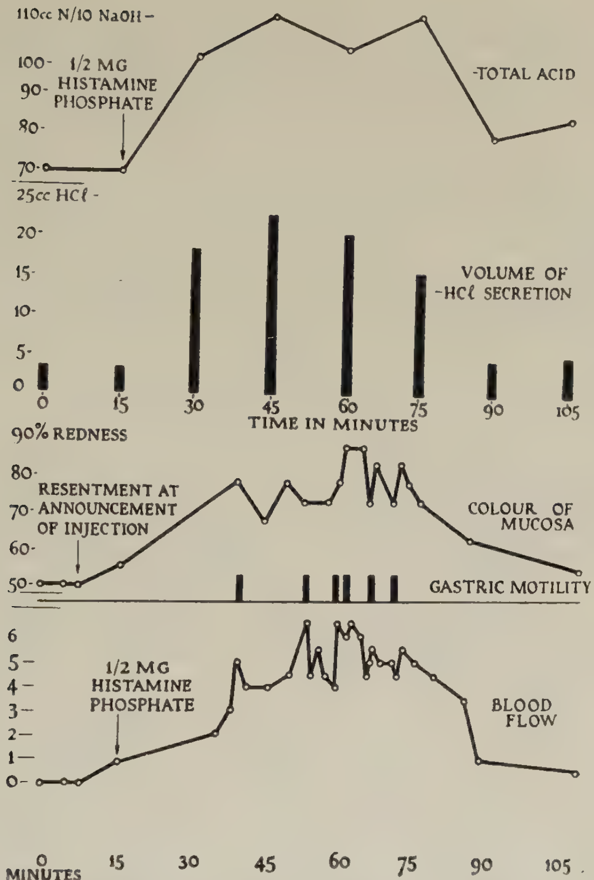


FIG. 5. 3. The effect of a hypodermic injection of histamine on gastric secretion, vascularity, motility and blood-flow in a human subject.

The vascularity was estimated as the redness, in arbitrary units, of the mucosa. The presence of large gastric contractions only is recorded. The blood-flow was measured by a thermal method, and is expressed in arbitrary units from an arbitrary zero. (Wolf and Wolff, " Human Gastric Function.")

consistency and composition, gastric emptying is usually completed in four to five hours.

While the food remains in the stomach it becomes mixed with the gastric juice, the secretion of which from the millions of tubular glands buried in the mucosa starts within a few minutes of eating.

Composition of Gastric Juice. The juice is really a mixture in variable proportions of the individual secretions of the various types of cell present in the glands. Heidenhain (1878) first recognised the *chief cells*, which contain pepsinogen, and the *parietal cells*, which secrete HCl. Both are absent from the pyloric region of the stomach ;

the glands there contain *mucoid cells*, which produce mucus, and between the pyloric region and the rest of the stomach there is a transitional zone where pyloric mucoid cells are mingled with chief and parietal cells. Besides the pyloric glands, mucus is secreted by cells in the necks of the glands elsewhere and also by the cells of the surface columnar epithelium.

The parietal cells are believed to secrete a fluid which is isotonic with blood, contains most of the water of the gastric juice, and is a practically pure solution of hydrochloric acid, the strength of which as secreted (160 milli-equivalents per litre) is constant whatever the rate of its formation. However, the acidity of the gastric juice is generally much lower than this maximal value, owing to neutralisation by the bicarbonate of the mucous secretion and buffering by the proteins, peptones and polypeptides of the gastric contents.

There is good evidence that the hydrogen ions of the gastric juice are derived by splitting water molecules, the hydroxyl ions remaining being neutralised by carbonic acid, with the formation of bicarbonate ions. The chloride ions of the gastric juice probably pass in from the blood passively, so as to preserve electrical neutrality. The carbonic acid is formed by the hydration of carbon dioxide, catalysed by the enzyme *carbonic anhydrase* (Chapter 3, p. 91), a large amount of which is found in the parietal cells.

The chief cells probably contribute a scanty non-acid secretion ; it contains *pepsinogen*, which is activated by acid, forming the proteolytic enzyme *pepsin*. The mucous cells produce a jelly-like fluid which contains much mucus and is faintly alkaline owing to the presence of bicarbonate.

Although stimulation of the vagus causes the secretion of a juice containing acid, enzyme and mucus indicating that the cells concerned all have a secretory innervation from the vagus, they can respond to some extent independently of one another to other forms of stimulation (mechanical or chemical) so that the final composition of the gastric juice may show wide variations. Thus the drug *histamine* is a powerful stimulant of the parietal cells, providing a juice of high acidity and containing little pepsin or mucus (its effect on the stomach of the subject " Tom " referred to in Fig. 5. 1, is shown in Fig. 5. 3) ; while mechanical or chemical irritation of the mucous membrane causes a profuse flow of mucus, with comparative little acid or pepsin. No selective stimulus for the chief cells is yet known.

Pepsin in acid solution breaks down proteins into peptones and proteoses, which are fairly large fractions of the original molecule ; some amino-acids are liberated, but the further breakdown of proteins and the above derivatives is accomplished later by the enzymes of the pancreatic and intestinal juices.

Stimulation of Gastric Secretion. Pavlov (1902) and his pupils were the first to show clearly that the secretion of gastric juice which starts within a few minutes of eating a meal occurs whether the food actually enters the stomach or not, and is due to a combination of " conditioned " and " unconditioned " reflexes similar to those causing the flow of

saliva under the same conditions. A dog was provided by a previous surgical operation with a gastric fistula for the collection of gastric juice, and an œsophageal fistula, so that the food which was swallowed never entered the stomach but fell out of the opening in the neck (Fig. 5. 4). A few minutes after the animal was thus " sham-fed " there began a flow of gastric juice, which could be stopped by cutting the gastric branches of the vagus, or paralysing them by the injection of the drug atropine. Sham-feeding was not always necessary to elicit secretion ; in intelligent animals, the mere sight, smell, or sounds associated with the arrival of food were sufficient.

These findings have been confirmed and extended by experiments on human subjects who have become accustomed by training to swallow and

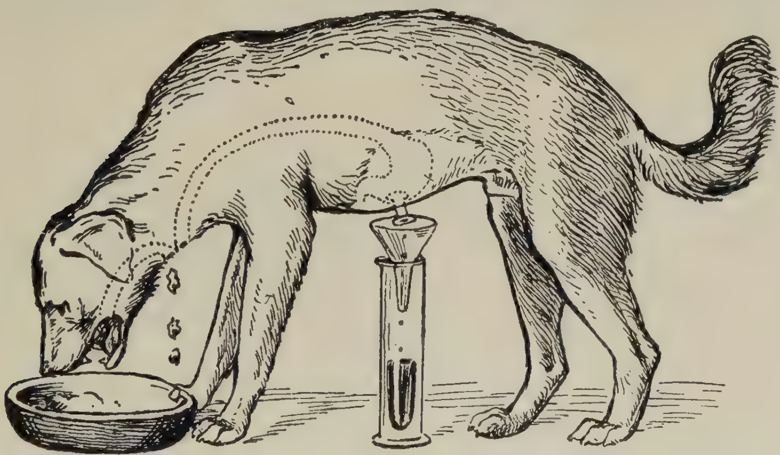


Fig. 5. 4. A dog with Œsophageal and Gastric Fistulæ.

The food consumed is seen dropping out of the open end of the anterior portion of the œsophagus ; the animal is fed through the opening in the posterior portion. The gastric fistula consists simply of a tube flanged at each end, stitched into the wall of the stomach at one end, and into the abdominal wall at the other. (Höber.)

retain without discomfort a stomach-tube for withdrawal of the gastric juice ; and occasional opportunities have also arisen of making similar and more extensive experiments on patients who, usually on account of an œsophageal stricture, have been provided by means of an operation with a gastric fistula for feeding. The classical example is that of Alexis St. Martin, an Indian " runner " at a trading station in Michigan, U.S.A., who was left as the result of a gunshot wound with a large gastric fistula. The observations and experiments made upon him (1825–33) by his physician, William Beaumont, have become famous. More elaborate studies of a similar kind have since been made by Carlson (1916) and by Wolf and Wolff (1943).

The reflex response to a " sham " meal gradually ceases in about an hour ; but if the swallowed food is allowed to enter the stomach in the usual way, to be digested by this juice (and later by the pancreatic and intestinal enzymes), gastric secretion is augmented and prolonged for three hours or more, in fact, long after the meal has been forgotten.

The cause of this continued secretion resulting from digestion of the food was first investigated by Pavlov by means of the famous " Pavlov pouch " (Fig. 5. 5). The secretion from this " miniature stomach " always runs closely parallel with that of the remainder, so that it becomes possible to follow the course of secretion in the main stomach during digestion, without interfering with it in any way.

Using dogs with such a pouch, and also a gastric fistula, Pavlov showed that placing meat in the animal's stomach without its knowledge (no reflex stimulation) caused little or no secretion, and digestion of the food took many hours ; but if the meat was first partly digested with " reflex " juice (obtained by sham-feeding another dog) the mixture stimulated gastric secretion strongly when introduced and digestion was rapidly completed. Not only does this experiment illustrate the importance of the " reflex " juice in gastric digestion, but

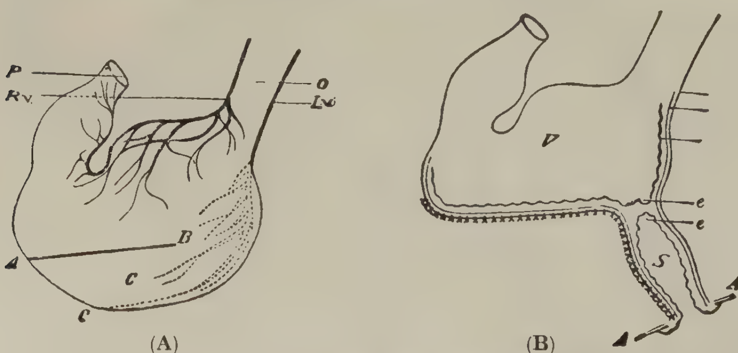


FIG. 5. 5. **The Pavlov Pouch.**

The left diagram (A) shows the line of the incision, *A—B*, into the gastric wall. *O*, œsophagus ; *R. v. L. v.*, right and left vagus nerves ; *P*, pylorus ; *C*, cardiac portion of stomach.

The right diagram (B) shows the operation completed (part sectional). *V*, main portions of stomach ; *S*, cardiac *cul-de-sac* (pouch) ; *A—A*, abdominal wall ; *e*, *e*, mucous membrane reflected to form a diaphragm between the two cavities.

it also indicates that digestion products are responsible for the later stimulation of gastric secretion.

Similarly, the introduction of food or its digestion products directly into the small intestine stimulates gastric secretion. The existence of this " intestinal phase " of gastric stimulation has been proved in dogs, by Ivy, by making the entire stomach into a pouch at an aseptic operation, joining the œsophagus directly to the duodenum. After recovery from the operation, when the animal eats a meal, this passes straight into the small intestine and is there digested ; a considerable secretion of gastric juice from the pouch occurs.

Gastrin. There is present in the mucosal cells of the antral region of the stomach a hormone named *gastrin* (Edkins, 1906) which is liberated into the circulation by the presence of food in the stomach ; it stimulates the parietal cells to secrete acid. The release of this hormone can be

demonstrated by providing a dog with two gastric pouches, one of the antrum and the other of the acid-secreting (fundic) region, and dividing all nervous connections between them. When the antral pouch is stimulated by distending it with a balloon or irrigating it with meat extract, the fundic pouch secretes. Gastrin has not yet been isolated, but potent and highly purified preparations which stimulate gastric secretion in the human subject have been made (Fig. 5. 6).

Experiments on conscious dogs have shown : (1) that the accumulation of acid in the antral contents during gastric digestion inhibits

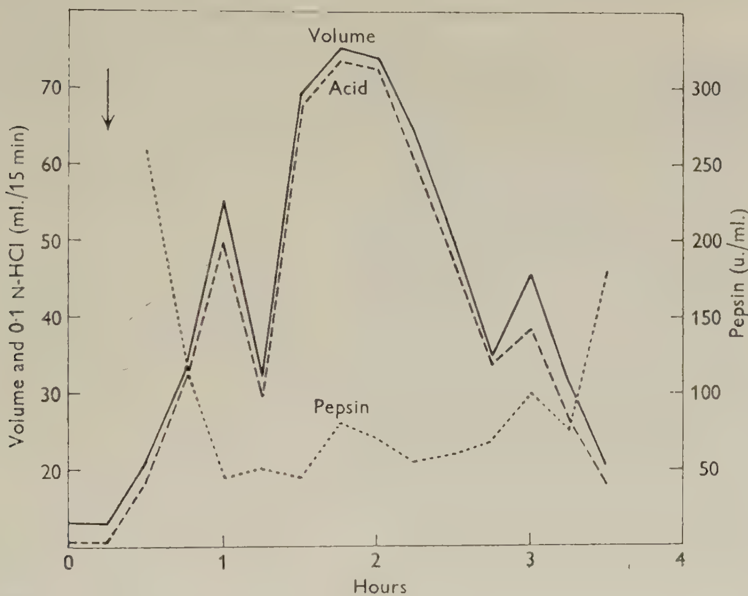


FIG. 5. 6. Response of the human subject to a subcutaneous injection of gastrin equivalent to 200 g. mucosa. (Gregory and Tracy, 1961.)

the release of gastrin, so that further secretion of acid is restrained ; and (2) that both the release of gastrin from the antrum by appropriate stimulation of this region, and the action of the hormone on the parietal cells, is potentiated by the concurrent reflex vagal excitation to both the antral and fundic regions of the stomach which occurs when a meal is eaten. The stimulation of acid secretion following a meal thus results from a combination of direct vagal excitation and the action of gastrin, on the parietal cells.

Histamine, a potent stimulus of parietal cell secretion, is present in all parts of the gastric and intestinal mucosa, particularly in the vicinity of the parietal cells ; it is also found in the gastric juice. The function of this histamine remains unknown ; it has been suggested that it plays some *rôle* in the activity of the parietal cells, but there is as yet no proof of this.

Gastric Inhibition. Besides the stimulation of gastric secretion from

the intestine, inhibition may be brought about by the presence there of fats, fatty acids and other substances. Since the last century, physicians have employed fat in the form of olive oil or cream, to depress gastric secretion in the condition of peptic ulcer, in order to promote healing. These substances are believed to liberate from the duodenal mucosa into the circulation a hormone *Enterogastrone*, which inhibits gastric secretion and motility. This is the physiological basis for the view that fatty meat such as pork is " indigestible " ; it takes a longer time to be digested and leave the stomach.

Movements of the Stomach

Some hours after a meal the normal human stomach is empty apart from a small and variable quantity of gastric juice, saliva, mucus, etc., and its walls are in a state of tonic contraction. When the swallowed food enters it, a *receptive relaxation* occurs as the result of a nervous reflex and the food slides down into the most dependent portion. Soon, as indicated in Fig. 5. 7, ring-like contractions appear in the body of the stomach and slowly move towards the pyloric sphincter, becoming deeper as they pass into this region where the muscle is stronger (W. B. Cannon, 1898). As digestion and emptying proceed, the strength and frequency of the contractions increase to a maximum which varies with the size and nature of the meal, and then gradually decline. Each wave occupies the stomach for about half a minute, and as many as four may be seen at the same time during the height of digestion. These contractions serve to mix the food with the gastric juice and, particularly in the pyloric antrum, provide the propulsive force for the passage of gastric contents at intervals into the duodenum.

Although the stomach relaxes when food enters, the tone is gradually regained, so that by the time most of the food has left the stomach a high tone is again present, with small regular fluctuations, termed a " tonus rhythm." This continues for a few hours after the stomach has emptied and then, if the next meal is not forthcoming, gives place to contractions similar in type to those normally seen in the filled stomach, but much more powerful. These occur in groups, lasting for about half an hour, and at intervals of two to three hours. Their incidence coincides with a sensation of hunger ; as they become stronger, definite pain—" pangs of hunger "—is felt with each contraction (Carlson, 1919).

Emptying of the Stomach. For many years it was believed that gastric emptying was chiefly controlled by the pyloric sphincter, which was supposed to remain closed for most of the time in the face of the gastric contractions, opening briefly at intervals to allow exit of some gastric contents. However, direct observations of the behaviour of the stomach and pyloric sphincter in human subjects by means of X-rays and the gastroscope, and experiments on trained conscious animals in which the regions concerned have been made accessible by the surgical preparation of fistulæ, show that the sphincter has no such independent *rôle*, but behaves like the pyloric antrum of which it is anatomically a part. In fact, the three regions, pyloric antrum, sphincter, and

duodenal cap act as a single co-ordinated physiological unit. As a gastric wave passes over each in turn, the antral contraction expels food through the still relaxed sphincter into the duodenal cap ; but this is brought to an end by the closely following contraction of the sphincter, and before the antrum and sphincter have relaxed, the contraction of the duodenal cap occurs, expelling the food down the duodenum. The effect is that of a "gastric pump" (Quigley, 1943), regurgitation from duodenum to stomach being prevented by the slightly persistent contraction of antrum and sphincter ; not every gastric wave results in this complete cycle of contractions, so that only

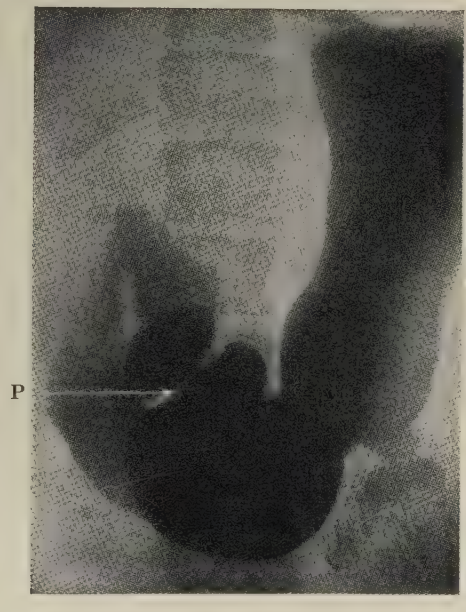


Fig. 5. 7. Radiograph of a human stomach after a "barium meal" showing peristaltic waves. P, pyloric sphincter. (F. Haenisch in A. E. Barclay, "The Digestive Tract").

a proportion of the waves which arrive at the pyloric antrum cause the exit of gastric contents.

Gastric emptying thus depends fundamentally upon the propulsive activity of the gastric muscle and the co-ordination of the three regions mainly concerned ; both these factors are controlled to a large extent from the duodenum. There are many substances besides fat, such as acid, hypertonic solutions and protein digestion-products, which retard gastric emptying when they are introduced into the duodenum ; this they do by causing reflex inhibition of the gastric musculature including the pyloric sphincter (*the enterogastric reflex*) as shown in Fig. 5. 8. In fact, by means of this reflex a constant restraining influence is normally exercised from the duodenum on gastric tone and motility ; if the reflex is prevented from operating during gastric emptying, gastric motility is greatly increased and the stomach empties abnormally rapidly (Fig. 5. 9).

Towards the end of gastric emptying, particularly of a fatty meal, the contraction cycles are weak, the pressures in antrum, sphincter and duodenum are nearly equal, and the sphincter is open most of the time ; such conditions are favourable for the regurgitation of intestinal juices and bile into the stomach, and evidence of this is afforded by the presence of these in samples of the gastric contents withdrawn by a stomach-tube.

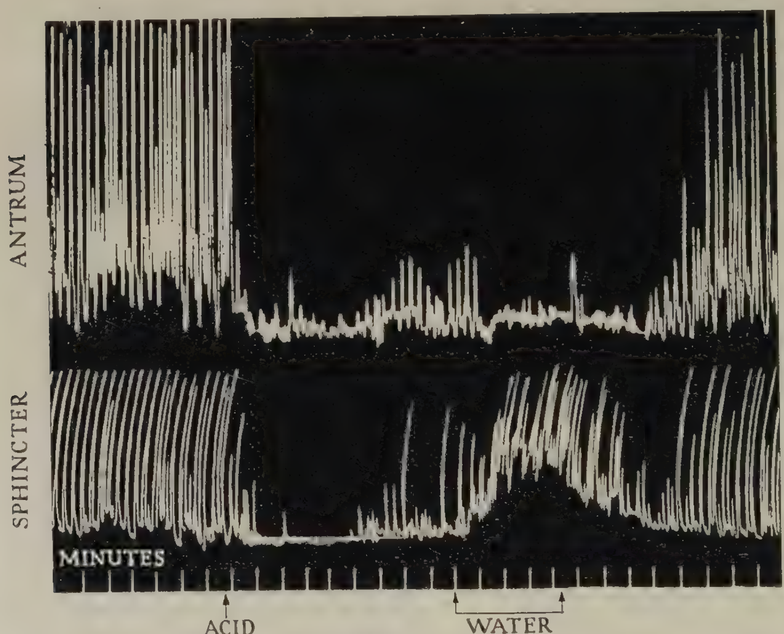


FIG. 5. 8. Record showing the effect on the pyloric antrum and sphincter in a conscious dog of : (*a*) injecting acid into the duodenum (20 ml. N/10 HCl) ; and (*b*) distending the duodenum with water (30 cm. pressure). Acid relaxes both the antrum and sphincter ; distension of the duodenum beyond the physiological range of pressure caused contractions and spasm (rise of baseline).

Gastric peristalsis was recorded by a balloon in the antrum, and contractions of the sphincter by a second balloon in the sphincter. (Thomas, Crider and Mogan.)

Vomiting. This is a reflex act involving the muscles of the diaphragm and abdominal wall and those of the stomach and œsophagus. It is co-ordinated by a " centre " in the medulla, which may be stimulated by irritation of any part of the digestive tract, by impulses from the semicircular canals (sea-sickness) or by disturbance of the centre itself (*e.g.* by cerebral tumours or the action of drugs such as apomorphine). A more or less prolonged sensation of *nausea* usually precedes retching and vomiting ; it is marked by pallor, sweating, salivation and partial or complete inhibition of the gastric musculature ; anti-peristalsis in the small intestine has been observed radiographically in human subjects. Nausea may culminate in *retching*, which consists of a series of inspiratory-like efforts accompanied by

closure of the glottis, the stomach becoming compressed between the diaphragm and the contracted abdominal muscles ; the gastric contents are finally ejected through the relaxed cardiac sphincter and œsophagus. The larynx is drawn up as in swallowing and elevation of the soft palate also occurs ; this largely prevents egress of the vomitus by the nose.

Fig. 5. 9. Record showing the effect of duodenal drainage on gastric peristalsis in a conscious dog provided with a high duodenal fistula (exclusion of enterogastric reflex). (Thomas, Crider and Mogan.)

Intestinal Digestion

As the stomach contents pass at intervals into the duodenum, they meet and mix with secretions from the pancreas, liver and intestinal glands, which complete the digestion of proteins, fats and carbohydrates, as the food passes down the intestine. The products are absorbed simultaneously into the portal and lymphatic circulations.

Secretion of Pancreatic Juice. The collection of pancreatic juice from a conscious dog by means of a cannula tied into the pancreatic duct, was first carried out by Regnier de Graaf (1664) and the method was revived nearly 200 years later by Claude Bernard, who gave the first description of the properties of the juice. Animals provided by a previous surgical operation with such pancreatic fistulæ remain in excellent health indefinitely, provided the juice is returned to the intestine daily, and not lost to the animal. For some purposes, however, collection of the juice for a few hours after cannulation of the duct in an anæsthetised animal is more suitable.

The acinar cells, which secrete the pancreatic juice, are apparently all of the same type ; they produce a secretion containing a number of enzymes and having a *p*H of 8–8·4 with a bicarbonate content which is approximately 1–5 times that in the blood, increasing with the rate of secretion. The amount

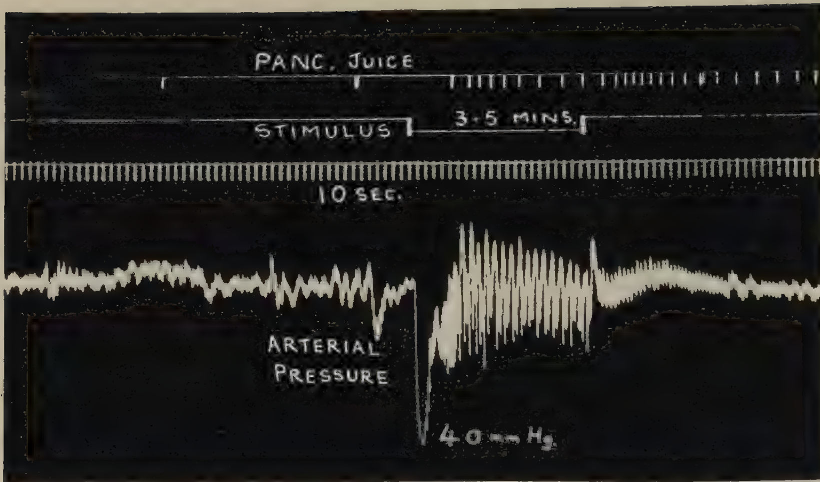


FIG. 5. 10. **Pancreatic secretion** produced in an anæsthetised dog by **stimulation of the vagus nerves** in the neck.

Note the effect on the heart-rate (cardio-inhibitory fibres in the vagus), the long latent period before secretion commences (forty seconds) and the scanty response. (Gregory, unpublished.)

FIG. 5. 11. **Pancreatic secretion** produced in an anæsthetised dog by the **injection of acid** (50 ml. of N/10 HCl) **into the duodenum.**

Note the copious and persistent secretion, compared with that given by vagus stimulation (Fig. 5. 10) in similar circumstances. (Gregory, unpublished.)

of chloride present is approximately inversely proportional to the rate of secretion, so that the sum of the concentrations of the two ions HCO_3^- and Cl^- remains about the same (Fig. 5. 12). The pancreas contains the enzyme *carbonic anhydrase* which is presumably concerned in the formation of the bicarbonate in the juice.

As in the case of the saliva and gastric juice, a reflex mechanism exists for the provision of pancreatic juice when a meal is eaten, and again it is the vagus which carries the secretory fibres to the gland (Fig. 5. 10); the secretion is scanty, but rich in enzymes. A much greater flow of juice occurs, however, as the gastric contents are passed on into the duodenum (Fig. 5. 11), and cause the liberation from the intestinal mucosa into the circulation of a hormone *secretin*, which

excites a copious and watery secretion from the pancreas. A second hormone *pancreozymin* is also released with secretin, and this, like excitation of the vagus, causes the secretion of pancreatic enzymes, with little or no effect on the volume rate of flow. These two hormones thus ensure the continued secretion of a copious flow of pancreatic juice, containing enzymes, during the intestinal digestion of a meal. The flow of hepatic bile and of juice from Brunner's glands in the duodenum is also excited to a small extent. The agent chiefly responsible for their liberation is the *acid* in the gastric contents, although fats, bile and protein digestion products (*e.g.* peptones) are also effective; acid is a much more powerful stimulus for the release of secretin than for that of pancreozymin.

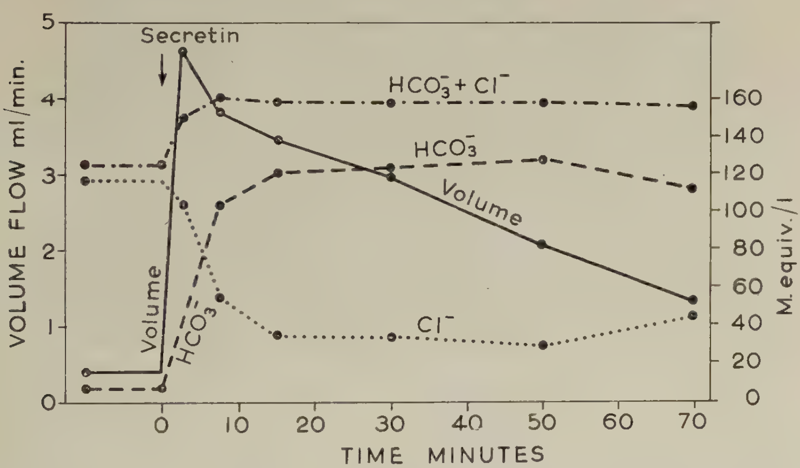


Fig. 5. 12. Human pancreatic secretion evoked by intravenous administration of secretin at time 0. The volume-rate and bicarbonate concentration are increased ; the chloride concentration is decreased. (Lagerlöf, 1942.)

Fig. 5. 12 shows the effect of an injection of secretin on the flow of pancreatic juice in a human subject ; the juice was withdrawn by means of a stomach tube which was swallowed and allowed to pass into the duodenum.

Secretin was the first hormone to be discovered. The fact that the entry of the acid gastric contents into the duodenum excited pancreatic secretion was well known to Pavlov and his contemporaries, but was ascribed to a reflex. However, in 1902, Bayliss and Starling showed that dilute acid still excited pancreatic secretion when placed in a *denervated* loop of small intestine, so that the effect must be mediated by way of the circulation. Intravenous injection of acid was without result ; but the injection of an acid extract of the intestinal mucosa caused a copious secretion of pancreatic juice ; and the active principle, secretin, has since been crystallised and identified as a polypeptide. Pure secretin does not excite the secretion of enzymes, although cruder concentrates of the hormone do so. In 1943, Harper and

Raper showed that such secretin extracts contain pancreozymin, now known to be a hormone, which stimulates the output of enzymes but has little action on that of bicarbonate, or on the volume rate of secretion. Preparations of both secretin and pancreozymin suitable for use in human subjects are now available commercially.

Actions of Pancreatic Juice. The pure juice is almost without action on most varieties of protein, the powerful proteolytic enzyme it contains being present in an inactive form *trypsinogen.* This is rapidly converted into the enzyme *trypsin* when the juice mixes with the intestinal juice which contains an enzyme-like activator *enterokinase.* Thus formed, trypsin acts upon all proteins and their digestion products, converting them finally into amino-acids and polypeptides. Pancreatic *diastase* breaks down starch into maltose, while the *lipase* also present hydrolyses the fats into fatty acids and glycerol.

The Bile. The entry of the gastric contents into the duodenum provides the stimulus for the appearance there of the bile, whose importance for digestion lies chiefly in the fact that the bile salts and lecithin it contains are valuable aids in the emulsification, digestion and absorption of the fats of a meal. It is also necessary for the efficient absorption of iron and of the fat-soluble vitamins.

Hepatic bile, a neutral golden-yellow slightly syrupy fluid, is secreted by the hepatic cells, and there is little evidence that its production is normally under nervous control. Between meals, the tone of the sphincter-like muscle around the duodenal end of the common bile duct is relatively high, and the bile flows into the relaxed gall-bladder where it is rapidly concentrated by the activity of the mucosa, becoming more viscid, very dark and slightly acid. When a meal is eaten, a little bile is sometimes reflexly expelled from the gall-bladder into the duodenum ; but the main emptying occurs later when contact with the duodenal mucosa of the gastric contents excites the liberation into the circulation of the hormone *Cholecystokinin* (Ivy), which causes slow contractions and emptying of the gall-bladder. Fat is particularly effective in liberating the hormone and hence in stimulating the emptying of the gall-bladder (Fig. 5. 13).

After the gall-bladder has emptied, hepatic bile may flow directly into the duodenum for a time until digestion there is over ; gradually the tone of the sphincter increases and the bile is once more diverted into the gall-bladder until the next meal.

Cholecystography. This clinical test of gall-bladder function depends on the fact that tetrabromphenolphthalein and similar compounds are opaque to X-rays and are excreted in the bile after oral or intravenous administration. They are concentrated in the gall-bladder and so enable it to be visualised by X-rays. If a meal rich in fat is then fed, the emptying of the gall-bladder may be recorded by serial radiographs (Fig. 5. 13).

The increase in the rate of flow of bile from the liver, which occurs during the digestion of a meal, is to some extent due to the passage through the liver of the products of digestion, but the chief stimulus is

supplied by the bile-salts themselves. Returning to the liver after absorption from the small intestine they stimulate the secretion of more bile, in which they are themselves incorporated. The total amount of bile salts thus circulating between liver, gall-bladder and intestine (*entero-hepatic circulation*), appears to be maintained at an approximately constant level by the liver ; if extra bile salts are administered to an animal, they are destroyed in a few days and the original circulating total is soon restored. If all the bile is drained from an animal having a biliary fistula and not returned to it, the rate of bile salt

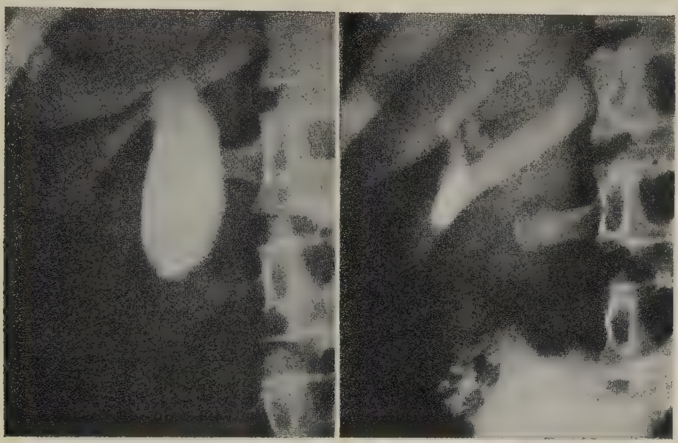


FIG. 5. 13. X-ray photographs of the **Gall-Bladder of a Man** before and after a meal of fat.

Tetra-iodo-phenolphthalein was injected intravenously fourteen hours before the first photograph was taken (Graham-Cole test). The second photograph was taken twenty minutes after the meal of fat. The discharge of the contents of the gall-bladder in response to the presence of fat in the duodenum has filled the cystic and common ducts with dye, and they can be seen, in the second photograph, forming a loop above the gall-bladder ; in some cases the hepatic duct becomes filled also. (Ivy.)

secretion falls to a low " basal " level ; it represents the rate of new bile salt formation by the liver, presumably in response to a maximal stimulus (the fall to zero of the quantity in circulation). When the return of bile is once more instituted, the quantity circulating rises in a few days to the original value.

The bile salts are derivatives of the steroid cholic acid, and are thus allied structurally to cholesterol and the sex and adrenal-cortical hormones (see Chapter 11). A number of different bile acids exist, but only a few are present in the bile of a particular species. A small amount only is present as the acid itself ; the rest is in the form of a compound of the bile acid with the base taurine or the amino-acid glycine. Taurocholic and glycocholic acids are present in human and ox bile : the dog, sheep and goat have only the former, the hog only the latter. How the liver synthesises the bile acids is still unknown.

The bile pigments. The hæmoglobin of worn-out red blood corpuscles is broken down by the cells of the reticulo-endothelial system, notably those of the liver (Kupffer cells), spleen and bone marrow, through the stages of hæmochromogens (Chapter 3, p. 72), which still contain the iron and globin of the original hæmoglobin molecule, to the *bile pigment biliverdin*, and its reduction product *bilirubin*, which are iron- and protein-free. The latter is set free into the blood, contributing to the yellow colour of normal plasma, and taken up from it by the liver, to be excreted in the bile. In some circumstances, *e.g.* in starving dogs, biliverdin is excreted by the liver in place of bilirubin.

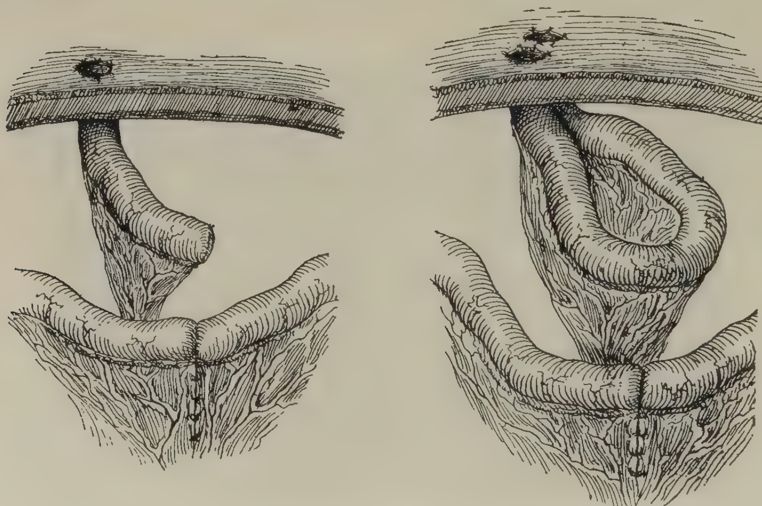


FIG. 5. 14. The Thiry and Thiry-Vella intestinal loops.
(From Markowitz, " Text-book of Experimental Surgery.")

The Intestinal Juices. The digestive juices contributed by the glands present in the wall of the small intestine come from *Brunner's glands* in the first inch or so of the duodenum, and *Lieberkuhn's glands*, which are found throughout the small and large intestines. Both secretions are alkaline and contain a good deal of mucus ; the stimulus for their appearance seems to be local mechanical and chemical excitation of the mucosa by digesting food.

The juice as ordinarily obtained—*e.g.* by distension with a balloon of an isolated loop of intestine (Thiry-Vella loop, Fig. 5. 14), contains small amounts of a variety of enzymes ; but similar enzyme activity is also demonstrable in extracts of the intestinal mucous membrane, and the invariable presence, in such samples of juice, of cast-off mucosal cells, leucocytes, etc., has given rise to the suspicion that most of its varied digestive properties may be due to *intracellular* enzymes liberated from the débris. If this material is rapidly removed from cat's intestinal juice by centrifuging it immediately after collection, the only enzymes found in appreciable amounts are lipase, amylase, and enterokinase. (Florey.)

The Movements of the Intestines

The mixing of the food with the intestinal secretions, and its passage through the alimentary canal, is accomplished by the intestinal move-

ments ; these are nicely co-ordinated with the progress of digestion and absorption so that both are virtually completed by the time the colon is reached.

A good way to gain some idea of the normal pattern of the intestinal movements (and incidentally, to study them experimentally), is to open under warm saline the abdomen of a decerebrate or lightly anæsthetised animal at the height of digestion. The inhibition caused by cold and drying is thus avoided ; and the movements of the coils of intestine, as they float outside the abdominal cavity, may be recorded by attaching them to levers writing on a smoked drum (enterograph), or by taking moving pictures which are analysed later ; or balloons may be inserted into the intestine and connected to volume or pressure recorders.

Many other methods have been used for study of the intestinal movements ; the more fruitful are probably those which utilise as a subject a conscious trained animal previously operated upon to render accessible the required region of the intestine (*e.g.* the Thiry-Vella loop).

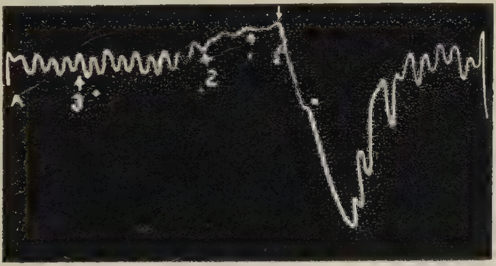


FIG. 5. 15. The passage of a bolus along the **Small Intestine.** Contractions of the longitudinal coat as recorded by an enterograph.

The bolus (of soap and cotton wool) was inserted into the intestine 4 in. above the recorded spot at the moment indicated by A. At subsequent moments it was 3 in., 2 in., 1 in. and ½ in. from the recorded spot as indicated below the tracing. As the bolus arrives 2 in. above the levers, there is cessation of the rhythmic contractions and inhibition of the tone of the muscle. This is followed, as the bolus is forced past, by a strong contraction on the rear of the bolus. (Bayliss and Starling.)

Three types of movement can often be distinguished. The " pendular movements " are a rhythmical lengthening and shortening of a segment of intestine and are probably caused by gentle waves of contraction which travel down the intestine for a short distance at about 2 to 5 cm. per second and occur about ten to twelve times a minute. Besides these one occasionally sees a portion of the gut which is the seat of a much stronger contraction of the circular muscle. This contraction obliterates the vessels and the lumen of the gut, blanching the intestine. It travels very slowly down the gut, about 0·1 to 0·5 cm. per minute, preceded by a less obvious but equally circumscribed region of inhibition. The double wave is known as a *peristaltic wave* (Bayliss and Starling, 1899), and is elicited by distending or stimulating the intestine strongly at any point. Such a response of the intestine (contraction

above, inhibition below, as shown in Fig. 5. 15), is probably due to a local reflex in the nerve-plexuses of the intestine, and has for this reason been termed the " Myenteric Reflex " (Cannon).

A third type of movement often seen is " segmentation " ; a portion of intestine, frequently the jejunum, becomes occupied by several simultaneous localised contractions. After a few seconds these disappear and are replaced by similar contractions in the intervening regions, so that the intestine is divided into a fresh set of segments (Fig. 5. 16). By this means the food is mixed with the digestive juices and brought into intimate contact with the mucous membrane.

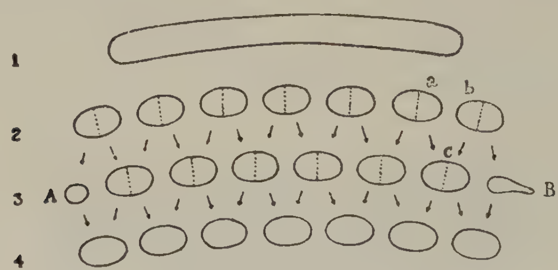


Fig. 5. 16. Segmentation in small intestine.

1. Loop of intestine before segmentation begins.
2. The loop is cut into little ovoid pieces by contraction of the circular muscle.
3. A moment later, each segment is divided into two parts, as shown by the dotted lines, and neighbouring parts, as a and b in line 2, run rapidly together and merge to form new segments, as c in line 3. The end pieces A and B are left small and move to and fro.
4. The process is repeated, with a return to the condition shown in line 2. (Cannon.)

The movements just described, and variations of them, are the groundwork from which is built up the normal complicated pattern of intestinal activity. But we do not yet understand very well how they are co-ordinated ; for instance, what determines the appearance of a given type of movement in some region of the gut, its intensity and range of influence, why it gives place to some other movement after a time, and finally how the movements as a whole are kept in step with the progress of digestion, so that the food moves along neither too quickly nor too slowly.

The intestine as a whole shows a descending gradient of activity throughout its length. After a meal, the duodenum and jejunum show great and varied activity ; but as the ileum is approached, the bowel becomes more and more quiescent, the terminal ileum making only occasional movements as it gradually fills with the residue of digestion.

Filling and Emptying of the Colon. As the stomach empties, the ileum is stimulated reflexly (*gastro-ileal reflex* of Hurst), to pass on its semi-fluid contents into the cæcum by sustained " stripping " contractions which occur every few minutes and persist while food remains

in the stomach. All movement of the cæcum is inhibited and it relaxes to receive the ileal contents. Gradually the cæcum, ascending and transverse colons are filled, without obvious peristalsis or other movement, the general appearance being one of " impressive immobility " (Hardy). There is a wide range of variation among normal persons in the time taken for different regions of the colon to become filled ; but as soon as cæcum and ascending colon are well filled, the saccular folds known as haustra appear (Fig. 5. 17A), and by their slow filling and emptying knead the contents and aid the absorption of salts and water.

FIG. 5. 17. (A) Radiograph of the human colon after a barium enema, showing haustra, and (B) a diagram (Holzknecht) of a " mass movement." (From A. E. Barclay, " The Digestive Tract.")

This slow and irregular process of filling is interrupted two or three times a day by a " mass movement " (Fig. 5. 17B). Starting usually about the middle of the colon, the haustra disappear and the colon becomes shortened and flattened by a rapidly advancing powerful contraction, and its contents are moved on bodily into the descending and pelvic colons in a few seconds usually without any subjective sensations whatever.

In most people the rectum is almost empty until just before the urge to defæcate or " call to stool " comes (commonly after breakfast) which is caused by a mass movement distending the rectum with fæces. The attainment in this way of a certain degree of distension initiates afferent impulses which are sent to a " defæcation centre " in the sacral region of the spinal cord. In newborn animals and infants, or after complete

transection of the spinal cord, efferent impulses then return from this centre, and produce contractions of the terminal colon, relaxation of the sphincters, and involuntary or " automatic " defæcation. In normal adults, filling of the rectum is appreciated in consciousness ; if circumstances are suitable the sacral centre is " permitted " and even " encouraged " by the cerebral cortex to operate as described above. Expulsion of the fæces is assisted by " straining," *i.e.* raising the intra-abdominal pressure by expiring against a closed glottis and contracting the abdominal muscles ; emptying of the anal canal is completed by contracting the levator ani muscles, which also restores the everted mucous membrane. If on the other hand defæcation would be inconvenient, activity of the sacral centre is inhibited by the cerebral cortex. The tone of the anal sphincters increases and that of the colon decreases ; the fæces in the rectum move back into the colon and the desire to defæcate disappears.

Colonic Secretion. The colonic mucosa contains very large numbers of mucous cells ; and Florey has shown that a secretion of mucus, accompanied by vasodilatation and contractions of the muscle, is produced by stimulation of the parasympathetic nerve supply, the pelvic visceral nerves. Stimulation of the sympathetic supply causes vasoconstriction and inhibition of movements without secretion.

Absorption of the Digestion Products

The Villi. The columnar epithelium of the small intestine is specially adapted for absorption of the products of digestion by the presence of the villi. These are finger-like projections of the surface, about 0·5 mm. long and containing a strand of muscle from the muscularis mucosæ, blood-vessels, nerves, and a central lymph vessel termed a lacteal ; the surface available for absorption is thus greatly increased. Between the villi open the mouths of the intestinal glands.

During fasting, the villi are shrunken and motionless ; but during digestion they swell up, due to the increased blood and lymph-flow through them, and contract rhythmically and independently of each other (Fig. 5. 18). These movements are probably of value in maintaining a good circulation through each villus and ensuring that this is constantly brought into contact with fresh portions of the intestinal contents. The water-soluble products of digestion, such as the amino-acids and sugars, are absorbed into the portal venous blood-stream and so pass through the liver before gaining the general circulation.

The pressure in the portal vein is about 20 mm. Hg, which is higher than the hydrostatic pressure of the intestinal contents. The intestinal wall, however, is impermeable to colloids, so that unless there is an appreciable colloid osmotic pressure within the intestine, water may well be absorbed as a result of the colloid osmotic pressure of the plasma proteins. The end result of the digestive processes is the breakdown of all colloidal material into crystalloidal ; finally, therefore, the intestinal contents exert no colloid osmotic pressure. But in the intermediate stages there may well be a substantial colloid concentration, and water may be drawn in from the blood.

Similarly, any substances which can diffuse through the intestinal wall will do so if their concentrations within the intestine are greater than those in the blood ; if there is no such concentration gradient in either direction, they will be carried through with the water. There is reason to believe that some substances of relatively small molecular weight normally leave the intestine in this way. On the other hand, other substances are transferred from the intestine to the blood proportionately more rapidly when in low concentration than in high by a " facilitated " process ; or they may be absorbed even when the intestinal concentration is less than the blood concentration, *i.e.* " up " the concentration gradient. Some secretory process, or " active transport," must be involved and metabolic energy, oxidative or glycolytic, is needed. As examples, we may mention urea, xylose and erythritol, which appear to be absorbed by diffusion only, while glucose and most amino-acids are absorbed, at least partly, by active transport. Sodium and chloride ions are absorbed rapidly, active transport (the " sodium pump ") being involved, but calcium, magnesium and sulphate ions are absorbed very slowly.

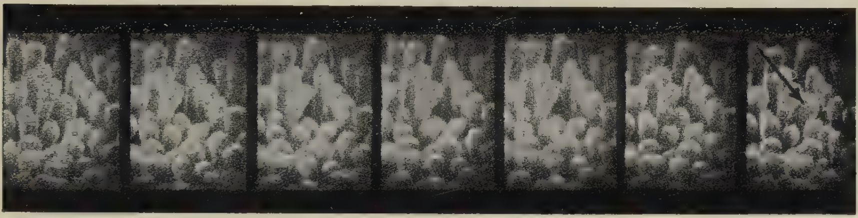


FIG. 5. 18. Portion of a cinematograph film following the **Movements of the Villi** in the living intestine of the dog.

The interval between each frame is approximately one second. Note that the villus indicated by the arrow in the right-hand frame becomes progressively shorter until it can only just be seen in the third and fourth frames from the right ; it then becomes longer again, and is only a little shorter in the last frame than it is in the first. (Kokas and Ludany.)

Such feats on the part of the intestinal epithelium are reminiscent of those performed by the kidney tubule cells in producing urine from the glomerular filtrate (Chapter 9) and the intracellular mechanisms involved are no doubt similar.

Absorption of Fat. During the absorption of a fatty meal, the lymphatics draining the intestine can be seen to be filled with a creamy fluid *chyle* which consists of lymph loaded with globules of neutral fat. This very early observation (Asellius, 1622) gave rise to the natural assumption that the emulsified fat in the intestines was absorbed unchanged, much as oil soaks through paper ; but Claude Bernard's discovery (1846) of the powerful lipase present in the pancreatic juice originated the view which is generally accepted to-day, that the greater part of the fat is hydrolysed in the small intestine before absorption ; the liberated fatty acids enter the mucosal cells and are there resynthesised into neutral fat, which passes into the central lacteal of the villus (Fig. 5. 19) and is ultimately discharged into the systemic venous blood via the thoracic duct. Most of the fat in the blood after a meal is in the form of minute droplets 0·5 to 1·0 μ in diameter, the *chylomicrons*,

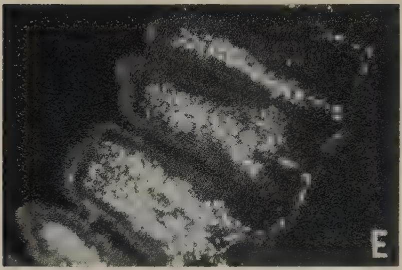


FIG. 5. 19. A photomicrograph of the small intestine of a rat, illuminated with ultra-violet light, during the absorption of fat containing vitamin A. This is fluorescent, and is visible in the epithelium, central lacteal of the villus and submucosal lymphatics. (Popper and Greenberg.)

which are believed to consist of neutral fat covered with a thin film of phospholipid (Fig. 5. 20).

Bernard also showed that the presence of the bile was necessary for the absorption of fat, as indicated by the appearance of chyle in the intestinal lymphatics ; the fatty acids liberated by lipolysis in the small intestine do not form soaps, the solution not being sufficiently alkaline, but are kept in solution and rendered absorbable by the bile salts.

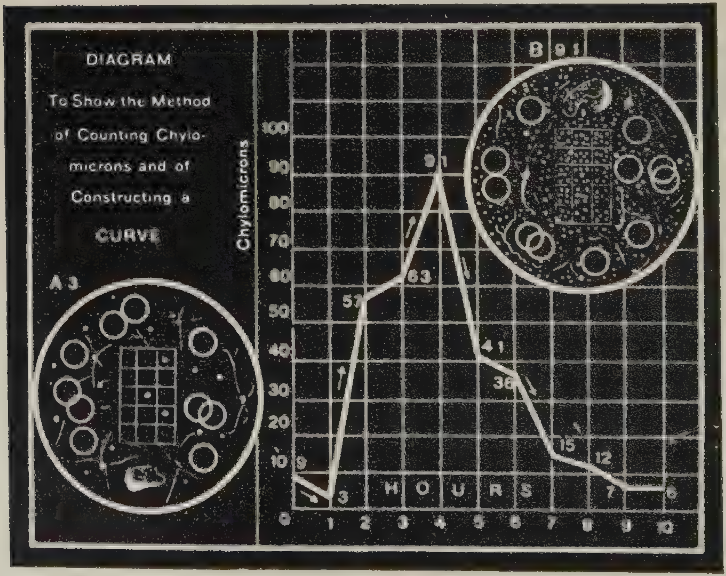


FIG. 5. 20. Curve showing the **Number of Chylomicrons** in a standard field, at various times after a meal of fat.

Two illustrative fields are also shown : A (on left) before absorption has begun, with 3 chylomicrons in the standard area ; and B (on right) with 91 chylomicrons. (From S. H. Gage and P. A. Fish.)

Besides the lymphatic pathway for the absorption of fat, there is an alternative pathway—the portal venous blood. The fatty acids of shorter chain length, which are more readily soluble in water, take this pathway to the liver. Thus, although almost all of the stearic (C_{18}) and palmitic (C_{16}) acids are found in the intestinal lymph after feeding, only about one-half of lauric (C_{12}) and one-fifth of decanoic (C_{10}) acids take this route, the remainder entering the portal blood chiefly as free fatty acids.

CHAPTER 6

METABOLISM

The tissues forming the animal body are composed of chemical substances all of which are derived from its environment, and during the animal's life these substances are returned to that environment many times over. In spite of this dynamic equilibrium the living animal maintains its individuality and indeed life is the continuity of this individuality. After death the continuity is broken, and the elements which composed the animal body become once more part of the environment. This maintenance of the individual depends on an elaborate series of chemical processes and the combination of all these reactions is called metabolism. These chemical reactions can be regarded mainly as of two kinds. The first is concerned with the building up of the complex substances of the body tissues out of simpler substances. These processes are responsible for the maintenance of the tissues by replacement of the loss due to wear and tear and in addition to this, in the young animal, they are responsible for growth and development. These synthesising reactions (sometimes grouped together under the term *anabolism*) do not supply the energy which the body needs for carrying out its functions ; indeed, they themselves require a certain amount of energy from some other source. The source of this energy, and also of the energy which the body needs for its other activities, is provided by a second set of chemical reactions which are grouped together under the term *catabolism.* In these reactions more complex chemical substances are broken down into simpler ones and this disintegration which is mostly of an oxidative nature, is accompanied by liberation of energy. The complex substances, capable of yielding energy on oxidation, are taken into the body as food and it has already been seen how the breaking-down process begins in the intestine. Digestion, however, is not an oxidative process, and is accompanied by liberation of only a very small fraction of the energy of the food substances. The main liberation of energy takes place after the food has been absorbed into the blood stream and it is to those changes subsequent to absorption that the term metabolism is generally applied.

The general aim of metabolic studies is to determine how the chemical energy of the food substances is utilised in contraction of muscles, secretion of glands, transmission of impulses along nerves, growth of tissues and the other activities characteristic of the living animal. The present position of the problem is that a very great deal is known about the chemical reactions which occur and the amount of energy made available, but relatively little is known about how this energy is used by the tissues for their purposes. Consequently most of this chapter is devoted to an account of the chemical processes

occurring in metabolism, the methods used for their study, the end products produced and the energy relations involved in these processes.

GENERAL METHODS OF METABOLIC STUDIES

The actual oxidative processes take place in the separate cells which compose the body, but the provision of the metabolic fuel, and the utilisation of the energy liberated, is only possible when the cells are organised into tissues, organs, whole animals and indeed societies of animals. On the other hand many of the chemical reactions cannot easily be studied in the whole animal, but only in isolated organs or tissues. Hence it is necessary to study metabolic problems at a series of different levels of biological organisation, *i.e.* to study the metabolism of individual cells, of isolated tissues, of organs and of whole animals. From the disconnected pieces of information obtained from these different sources we try to form a composite picture of the whole process of metabolism. Each type of metabolic experiment has its own special use. In the case of the whole animal we can administer substances by mouth, intravenously, intraperitoneally or subcutaneously, collect the waste products in the urine, estimate chemically the changes in the blood, measure the gaseous metabolism and thus study the energy relationships under physiological conditions. In particular, studies on the whole animal are essential for investigation of the effect of hormones on metabolism and the total nutritional requirement of the individual. Under this heading must also be placed the mass of clinical observation which has contributed greatly to our knowledge of physiological processes in the field of metabolism. Experiments with various organs make possible a rather fuller study of the chemical changes undergone by the food substances and also bring the oxidative processes into relationship with the special activity of each particular organ. The use of isolated tissues has made possible a very detailed study of the enzyme systems of the body and of the chemical changes brought about by these, even although the tissues in the experiments are not working under physiological conditions. The study of the individual cells has shown that at least some of the enzymes can be located in definite cells and in some cases in definite positions in the cell.

The Respiratory Quotient

Whatever method we use to investigate metabolism we consider the amount of oxygen used, the amount of carbon dioxide formed, the nature and amount of the foodstuff oxidised and the energy liberated, and all these must bear a strict relationship to each other. Since of these quantities, the amounts of carbon dioxide and of oxygen are often the simplest to measure, and also because they give easily obtainable quantitative information about metabolism, it is of great importance to understand their full significance.

If the food substance being oxidised is glucose, the reaction (or rather the sum total of a large number of intermediate reactions) is as follows:

$$C_6H_{12}O_6 + 6O_2 = 6CO_2 + 6H_2O$$

Since the volumes of gases are proportional to the numbers of molecules present, the volume of carbon dioxide formed will be exactly equal to the volume of oxygen used. The ratio of carbon dioxide produced to oxygen used, the **respiratory quotient**, in this case is equal to one. In a similar way if equations for the oxidation of fats are written down the respiratory quotient can be calculated for each case; thus for triolein

$$C_{57}H_{104}O_6 + 80\ O_2 \rightarrow 57\ CO_2 + 52\ H_2O$$

The respiratory quotient in this case is 57/80 or 0·71. In the case of proteins, it is not possible to work out the respiratory quotient in this simple way, but from a knowledge of the percentage composition of the protein, and the products of oxidation in the body, it is possible to calculate that the respiratory quotient is about 0·8.

The respiratory quotient for all forms of carbohydrate is 1·0. The exact value of the respiratory quotient will vary for different fats and proteins, but by taking a mean value for those normally present in the food it can be determined that the respiratory quotient for a diet consisting of fat would be 0·71 and for protein 0·81. A knowledge of the respiratory quotient therefore gives us information about the type of food which is actually being oxidised, and since the energy liberated by a certain amount of oxygen depends on the nature of the substance oxidised, it is evident that, from the amount of oxygen used and from the respiratory quotient, much information can be obtained about metabolic activities. The significance of the respiratory quotient will be considered more fully later in connection with the metabolism of the whole animal.

Metabolism of the Cells

The study of the metabolism of the living cell in relation to its structure is called histochemistry or cytochemistry, and the methods used are largely a combination of histological and biochemical techniques. By disintegration of cells and subsequent centrifuging it has been possible to divide the cell contents into different fractions, and many attempts are being made to determine the enzyme content of different cell fractions. Although some success has been achieved—*e.g.* the enzymes concerned with the oxidative processes in the cells have been mainly located in the mitochondria—it is not yet possible to go very far in analysing the metabolic functions of different parts of the cell.

Metabolism of Isolated Tissues

In the higher animals respiration is often divided into external respiration and internal respiration. External respiration includes all the processes which result in oxygen being brought to the tissues and carbon dioxide being removed. In contrast to this, internal respiration includes the processes which take place inside the cells of the body resulting in the oxidation of the food substances and the formation of the waste products. The study of internal respiration is largely carried

out by investigating the chemical changes which take place in isolated tissues. The remarkable thing is that most living tissues, disintegrated to a greater or lesser degree by mincing or slicing and deprived of their blood supply, can still, under suitable conditions, take up oxygen, carry out chemical reactions and produce carbon dioxide. The requisites for respiration in isolated tissues are a supply of available oxygen, the maintenance of normal body temperature, and the provision of a suitable fluid medium in which the tissue is suspended. The media usually employed are balanced salt solutions (see Chapter 8) and the other essential conditions are achieved by the use of microrespirometers, the usual type being that developed by Warburg.

The tissue, either minced or sliced, is put into a glass vessel which can be attached to a manometer. The manometer is fixed on a stand so that it, together with the attached cup, can be shaken continuously with the cup immersed in a water bath at constant temperature. A centre compartment in the cup contains a small piece of filter paper soaked in caustic soda, and this absorbs the carbon dioxide produced. After introducing the tissue suspended in a suitable saline medium, the whole apparatus is filled with oxygen, and the shaking apparatus is then set in motion. Readings are taken at regular intervals and the consumption of oxygen is indicated by the change of pressure in the manometer. The manometer cups are usually provided with one or two side arms into which reagents can be put which are to be added during the course of the experiment. Analysis of the contents of the cups can also be made at the end of the experiment and the products of metabolism estimated. There are many modifications of the manometers and cups and it is possible by various techniques to measure the respiratory quotient or the changes taking place anaerobically.

It is important to appreciate both the possibilities and the limitations of experiments with minced and sliced tissues. There is no doubt that the conditions are highly unphysiological, the tissue is deprived of its blood supply and its normal relations with other tissues. The chemical reactions which can be demonstrated by these techniques do not necessarily take place in the tissues under more physiological conditions and furthermore, we derive no information about how the tissues utilise the energy liberated by the reactions which occur. At the same time the method is invaluable for studying the various enzyme systems present in the cells, and for finding out what substances actually take part in the chemical reactions in the tissues. It is not too much to say that the enormous advances in many fields of biochemistry have been due in great part to experiments of this kind.

From the study of minced and sliced tissues it has been found that the food substances oxidised do not combine directly with oxygen to form the final products of oxidation, but pass through a number of intermediary stages. One of the methods of studying these intermediary stages is to interfere with the normal process of metabolism by blocking the series of reactions at some point so that intermediary substances, which would normally be further oxidised, accumulate in sufficient amounts to allow their detection and estimation. A number of such blocking agents or inhibitors is known, which will specifically put out of action certain enzymes in the tissues, while allowing others to function

normally, and by the use of these many of the intermediate products of metabolism have been identified. A few of the outstanding results of this kind of work are referred to in later sections, but for a more complete account of the chemical changes undergone by the food substances, the reader is referred to textbooks of biochemistry.

Metabolism of Isolated Organs

It is often useful to study the metabolism of some particular organ in relation to its functional activity, *e.g.* to study kidney metabolism in relation to the formation of urine or cardiac metabolism in relation to the pumping activity of the heart. For such purposes it is necessary to use whole organs instead of isolated tissues, but it is also necessary to separate in some way the metabolism of the organ being investigated from the metabolism of the animal as a whole.

Perfusion Experiments. One method of attacking such a problem is to supply the organ with nutrient material and oxygen by an artificial perfusing system instead of by its own circulation. Defibrinated blood or heparinised blood is usually employed for the purpose and this is saturated with oxygen, either by passing it through the lungs or through an oxygenator, and is then pumped through the vessels of the organ to be studied. By such a system organs can be kept " alive " for a number of hours after complete isolation from the rest of the body. The metabolism of heart, lungs, liver, kidney, brain and limbs have all been studied in this way. The gaseous metabolism is measured by estimating the volume of oxygen taken up by the blood and the volume of carbon dioxide given off, or alternatively it may be measured by estimating the content of oxygen and carbon dioxide in the blood supplying and leaving the organ and also the rate of blood-flow. In some cases, *e.g.* heart and voluntary muscle, the physical work done can be measured and compared with the rate of metabolism ; in other cases, *e.g.* the kidney, the osmotic work can be studied. In some organs, however, such as the brain, we know very little about the quantitative relation between functional activity and metabolism.

Metabolism of Organs in situ. Another method of studying the metabolism of individual organs is to leave the organ *in situ*, and measure the oxygen and carbon dioxide content of the arterial and venous blood together with the rate of blood-flow. For this purpose it is very useful to have an apparatus which will register continuously the amount of oxygen in the arterial and venous blood.

This can be done by means of two *oximeters* (Chapter 3, p. 74) in which the degree of oxygenation of the blood is measured by means of photo-electric cells, in terms of the optical transmittance at certain appropriate wave-lengths of the light. A continuous graphic record may thus be obtained of the oxygen content of the arterial and of the venous blood.

The Metabolism of the Intact Animal

When the metabolism of the whole animal is considered it is usually not in terms of the intermediate products of metabolism but rather of

the sum total of the metabolic reactions, *i.e.* the amount of food and of oxygen used and the amount of carbon dioxide and waste products formed. It is also possible to measure the total heat production of the animal and the amount of physical work carried out, so that balance sheets can be prepared of the total intake and output of energy in all its various forms, thermal, chemical, mechanical, etc. Very careful measurements of these energy relations have shown that the animal body behaves exactly like all other chemical or mechanical systems as regards the law of the conservation of energy, in that the total energy produced is equal to the total energy supplied. The vital activities of the body are in no way a creation of energy but simply a transference of energy from one form to another. The particularly " vital " part of the process is that some of these energy transformations can only take place in living tissues and so far have not been imitated in non-biological systems.

The Energy Value of the Foods. In order to prepare our complete balance sheet we require to know the energy values of the foods taken. For this purpose we assume that oxygen is freely available and therefore consider the amount of energy capable of being liberated by oxidation of the food. This can be determined outside the body by means of the bomb calorimeter.

This consists of a strong steel chamber which can be sealed by a tightly fitting lid. Into the chamber a measured quantity of the food substance is introduced and the whole apparatus filled with oxygen at high pressure. The bomb calorimeter is now placed in a known volume of water at a certain temperature. Combustion of the contained substance is initiated electrically and the amount of heat produced is estimated from measurement of the rise in temperature of the surrounding water.

By means of the bomb calorimeter it is found that 1 g. of carbohydrate produces 4·1 kilocalories, 1 g. of fat 9·2 kilocalories and 1 g. of protein 5·3 kilocalories. These are the energy values when combustion is complete. In the body complete combustion of carbohydrate and fat takes place, but in the case of protein the end product, urea, is still capable of further oxidation, though not inside the body, and hence the energy value of the urea must be subtracted from that of the protein. Making these corrections the values for the three types of food are : carbohydrate and protein each 4·1 kilocalories per g. and fat 9·2 kilocalories per g. Knowing these values and also the amount of each food substance in the diet we can calculate the total energy provided by the food. This knowledge of the energy content of food is essential in making up diets for people under various conditions.

Energy Production from Oxygen Intake. The food taken into the body gives the total energy intake but it does not tell us the rate of metabolism at any one time, since the food is not all used immediately but may in part be stored. The rate of metabolism is derived from the oxygen intake, for we know the oxygen intake only keeps pace with the immediate metabolic needs. If we wish to know the metabolic rate approximately, we can calculate it from the oxygen consumption by

assuming that each litre of oxygen used in the body yields 4·8 kilocalories. Thus in the human subject where there is an oxygen consumption of 300 ml./min., the energy production would be 87 kilocalories per hour. This degree of accuracy is in fact sufficient for many purposes.

If we require to know the metabolic rate more accurately, we must consider the fact that the energy liberated by the consumption of 1 litre of oxygen in the body is not a constant figure of 4·8 kilocalories, but varies with the kind of food being oxidised. The food taken is a mixture of protein, fat and carbohydrate, so that in addition to determining the oxygen consumption we have to determine what proportions of these three kinds of food are present in the diet. This can be done in the following way.

The total nitrogen in the urine is estimated over a known time. By multiplying this value by 6·25 one can calculate the total amount of protein metabolised during that period. (The average amount of nitrogen in dietary proteins is 16 per cent.) We can further calculate the amount of oxygen needed to oxidise this protein in the body and the amount of carbon dioxide produced. These amounts of oxygen and carbon dioxide are now subtracted from the total amounts of the oxygen and carbon dioxide involved in metabolism, and the resultant figures give the oxygen used and the carbon dioxide produced in non-protein metabolism. Since there are now only two substances to be dealt with, fat and carbohydrate, and since we know the respiratory quotient corresponding to each, we can calculate the proportions of each necessary to give the respiratory quotient of the non-protein metabolism. In practice one gets the result from tables already worked out for each possible respiratory quotient (Table 6. 1). Such tables also give the energy production per litre of oxygen for each respiratory quotient so that the total energy production can readily be calculated once the total oxygen consumption and the non-protein respiratory quotient are known.

TABLE 6. 1

The relation between the Respiratory Quotient, the relative amounts of fat and carbohydrate oxidised, and the energy production of the non-protein metabolism.

Non-protein Respiratory Quotient		1·00	0·95	0·90	0·85	0·80	0·75	0·718
Per cent total O_2 consumed by carbohydrate		100	82	65	47	29	11	0
Grams Foodstuff per litre O_2	Carbohydrate	1·23	1·01	0·80	0·58	0·36	0·14	0
	Fat	0	0·09	0·18	0·27	0·36	0·45	0·50
	Total	1·23	1·10	0·98	0·83	0·72	0·59	0·50
Kilocalories per litre O_2		5·05	4·99	4·94	4·88	4·83	4·77	4·74

Since the respiratory quotient for protein is intermediate between that for fat and for carbohydrate it is often considered sufficiently accurate to neglect altogether protein metabolism, and treating the whole metabolism of the animal as non-protein metabolism, to make the calculations accordingly.

In making calculations from the respiratory quotient it is assumed that the amount of carbon dioxide given out in the expired air is the same as that being produced by the tissues. While this is true if the carbon dioxide output is considered over long periods, it is not necessarily so over short periods. During strenuous exercise lactic acid accumulates in the blood and this results in liberation of some of the carbon dioxide stored in the blood as bicarbonate, so that the amount of carbon dioxide expired by the lungs is greater than the amount which is being formed by the tissues. In such conditions, the respiratory quotient can in fact exceed 1·0. Similarly, in the period of recovery after exercise when lactic acid is being removed from the blood, the amount of carbon dioxide expired may be less than the amount formed by the tissues, as some is being retained by the blood to form bicarbonate. Another factor which can contribute to changes in the respiratory quotient is interconversion in the body of the various food substances (p. 192).

Methods of Determining Metabolic Rate in the Intact Animal

Two different principles are used for this purpose—direct calorimetry and indirect calorimetry. In the former the object is to measure the heat production directly, while in the latter the heat production is calculated from the gaseous exchange.

Direct Calorimetry. This involves very elaborate and expensive apparatus and is therefore little used in routine measurement. For some purposes, however, it gives information not obtainable in any other way, *e.g.* if we wish to draw up a complete balance sheet of the energy exchanges in the body. The subject is placed inside a large calorimeter, which in the case of the human subject is a small thermally-insulated room known as the Atwater-Benedict Respiration Calorimeter (Fig. 6. 1). The calorimeter is in fact constructed so that the measurement of heat production is combined with the indirect method of measuring metabolism, *i.e.* by the gaseous exchanges ; and these accessories are shown in the diagram. An energy balance drawn up from procedures of this kind is shown in Table 6. 2.

Indirect Calorimetry. The principle of calculating the metabolic rate from the oxygen consumption has been discussed above. The problem at the moment is the technique of measuring the gaseous exchanges. Two different principles are used for this purpose and these are distinguished as closed methods and open methods.

Closed Methods. These in general require a less mobile equipment and hence are limited in their use. In the simpler types of apparatus for the human subject the expired air, collected from a mouthpiece supplied with valves, passes over soda lime which absorbs the carbon dioxide. The soda lime is usually contained in a recording spirometer,

and from this the subject rebreathes his own expired air freed from carbon dioxide. During this process the volume of the air in the circuit diminishes at a rate equal to the consumption of oxygen and by recording the volume change by means of the spirometer the rate of oxygen

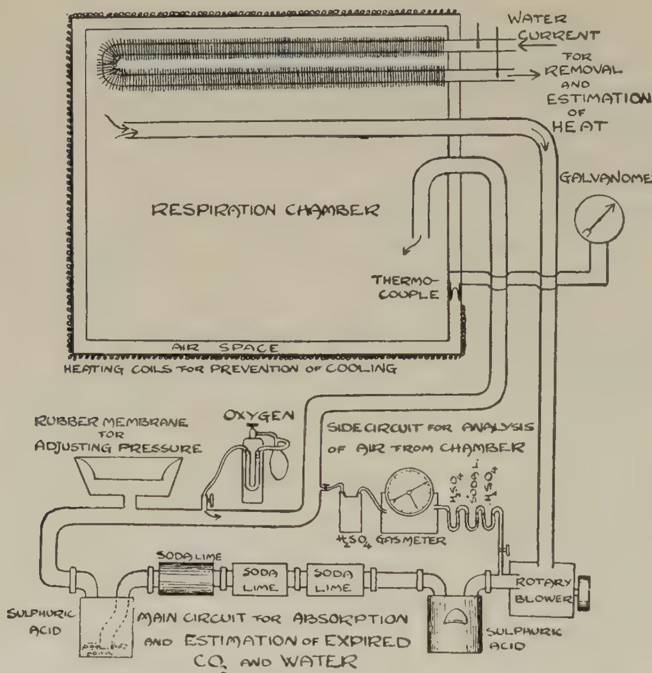


Fig. 6. 1. Diagram to illustrate the principle of the **Atwater-Benedict Respiration Calorimeter.**

The upper part represents the calorimeter, in which the subject is placed, and his heat production measured in terms of the rise in temperature and rate of flow of the cooling water (in practice the cooling pipes are carried right round the chamber).

The lower part represents the **Benedict Respiration Apparatus,** in which the carbon dioxide produced by the subject is absorbed by the soda-lime, and the oxygen used is replaced from the cylinder through a meter (not shown) at such a rate as to keep the rubber membrane of the pressure equaliser in a constant position. The gas is dried in the sulphuric acid bottles before entering the carbon dioxide absorbers, since soda-lime absorbs water as well as carbon dioxide. Any water that may be evolved from the soda-lime in its reaction with carbon dioxide is absorbed in the second sulphuric acid bottle.

When the respiration apparatus is used for indirect calorimetry only, the tubes leading to the calorimeter are connected to a mouth-piece, which is held between the subject's teeth, his nose being closed by a clip. (Parson's "Fundamentals of Biochemistry.")

consumption can be measured. The apparatus is illustrated in Fig. 4. 7 (p. 104). It will be noted that in such a method no account is taken of the respiratory quotient and the metabolism is calculated by assuming an average value for the respiratory quotient.

TABLE 6. 2

Energy Balance Sheet for Human Subject.

I.—Indirect Calorimetry

Food Ingested.[1]		Class of food.	Food Oxidised.[2]	
Weight, g.	Energy, kcal.		Weight, g.	Energy, kcal.
79·2 59·6 201·0	448 569 842	Protein Fat Carbohydrate	64·8 117·8 226·3	366 1,124 948
339·8	1,859	Total	408·9	2,438
69·1	579	From body stores (*i.e.* excess of food oxidised over food ingested).		
408·9	2,438	Total	408·9	2,438

II.—Direct Calorimetry

Heat produced, as measured	2,334 kcal.
Potential energy of urine	90 ,,
Total heat produced	2,424 ,,

III.—Test of Law of Conservation of Energy

Heat production as calculated from results of indirect calorimetry	2,438 kcal.
Heat production as observed by direct calorimetry	2,424 ,,
Difference . . .	14 ,,
	i.e. 0·6 per cent.

[1] Corrected for losses in digestion and absorption.
[2] Calculated from respiratory exchange.

By means of more elaborate circuits it is possible to measure the carbon dioxide production as well. In such cases the air is circulated by a pump and passes through containers with sulphuric acid to absorb moisture, and with soda lime to absorb carbon dioxide. These can be weighed periodically and hence the production of water and carbon dioxide measured. In these systems, the oxygen consumption is measured by adding oxygen at a known rate so as to replace the amount used. This is most conveniently done by having in the circuit, at some point, a sensitive rubber membrane, which will indicate alterations in the total pressure and hence in the volume of the system. Such circuits are often used for small animals. For application to man, Benedict's apparatus is employed. For more complete metabolic estimations in

man, this system may be combined with the Atwater-Benedict Respiration Calorimeter.

By means of such apparatus a complete balance sheet of the energy exchanges in man can be made, and since much can be learned from the study of this, the results of one such experiment are given in Table 6. 2.

Open Methods. In these the subject breathes in from the atmosphere, but by means of a mouthpiece containing valves the expired air is collected in a Douglas bag (Chapter 4, p. 105). The volume of air expired in a given time is measured and a sample is taken for determination of the percentage of oxygen, nitrogen and CO_2 present. The oxygen utilisation is obtained by calculating the amount of oxygen in the expired air in a given time and subtracting this from the amount of oxygen in the inspired air in the same time. Similarly, the CO_2 produced is the amount of CO_2 in the expired air from which has been subtracted the amount of CO_2 in the inspired air (an almost negligible quantity). In making these calculations, particularly for oxygen, it must be remembered that the volume of inspired air is not the same as the volume of expired air, owing to the fact that the volume of oxygen taken in is greater than the volume of CO_2 given out, unless the R.Q. is exactly 1. It is the volume of expired air which is measured, but since the *amount*, although not the percentage of N_2 in the inspired and expired air, is the same, the volume of inspired air can be calculated as follows :

$$\text{Vol. of inspired air} = \text{Vol. of expired air} \times \frac{\text{Percent. } N_2 \text{ in expired air}}{\text{Percent. } N_2 \text{ in inspired air}}$$

The open methods of measuring gaseous exchanges have the merit of great mobility, and this enables metabolic rate to be measured in a great variety of conditions. In recent years much improved forms of the open method have been devised, in which the whole apparatus is more portable and very little resistance is applied to the flow of expired air, and one of the best of these is Wolff's Integrating Motor Pneumotachograph.

The Basal Metabolic Rate

The rate of metabolism depends on the amount of physical work which the body does and, therefore, can be reduced if the subject remains completely at rest. But even when all unnecessary movements are stopped a certain amount of energy is used in maintaining the body temperature and in providing for the needs of circulation, respiration and other vegetative processes, and this cannot be further reduced without damage to the tissues. This amount of metabolism is considered the basal level, on which extra metabolic activities are superimposed when the body undertakes more work. It is called the basal metabolic rate, or more usually the B.M.R. One of the functions of this basal metabolism in warm blooded animals is to keep the body temperature above that of its surroundings and, since the loss of heat from the body takes place from the surface exposed to the atmosphere, it is not surprising to find that the B.M.R. is more closely related

Table 6. 3

The Basal Metabolic Rates of Various Animals.

Animal	Weight in kg.	Kilocalories produced per day	
		per kg. of weight.	per sq. m. of surface.
Horse . . .	441	11·3	948
Pig . . .	128	19·1	1,078
Man . . .	64·3	32·1	1,042
Dog . . .	15·2	51·5	1,039
Mouse . . .	0·018	212	1,188

to the body surface than to the body weight. This is well illustrated in Table 6. 3, which gives the B.M.R. for animals of different sizes expressed per unit weight and per unit body surface.

The area of the body surface is not an easy quantity to measure directly. In man it is usually obtained from the height and weight by means of the following formula of Du Bois,

$$S = 0{\cdot}007184 \times W^{0{\cdot}425} \times H^{0{\cdot}725}$$

where S is the body surface in sq. metres, W the weight in kilograms and H the height in centimetres. It can be obtained more easily from nomograms based on this formula (Fig. 6. 2).

In a well-nourished man the B.M.R. has been found to be about 1,700 kilocalories per day. This is equivalent to about 40 kilocalories per square metre body surface per hour or to about 1 kilocalorie per kg. body weight per hour. It varies with sex and with age, being higher in males and young people. It is usual to express the B.M.R. of an individual as a percentage increase or decrease above or below these standard values. In normal individuals, the values lie within about 15 per cent. of the standard values.

Since the metabolic activities of the body can be increased by taking food or by performance of work, it is very important in measuring the B.M.R. that the subject should be at complete mental and physical rest and should be in a condition of fasting for about twelve hours. Extremes of temperature, previous diet, previous exercise, emotion or menstruation may all have some effect on the result.

The Thyroid Gland and the Basal Metabolic Rate. Of the pathological causes of alteration in the B.M.R. the commonest is abnormal activity of the thyroid gland discussed in Chapter 11. In diseases where there is thyroid deficiency, as in cretinism and myxœdema, there is a low B.M.R. and this can be increased by administration of thyroid extract. In exophthalmic goitre there is hyperthyroidism and here the B.M.R. is considerably increased. Surgical removal of part of the thyroid gland leads to a fall in the B.M.R. along with alleviation of the other symptoms. The active principle in thyroid extract which is responsible for the effect on metabolism is thyroxine. The stimulating effect of this and related compounds on metabolism has been used for

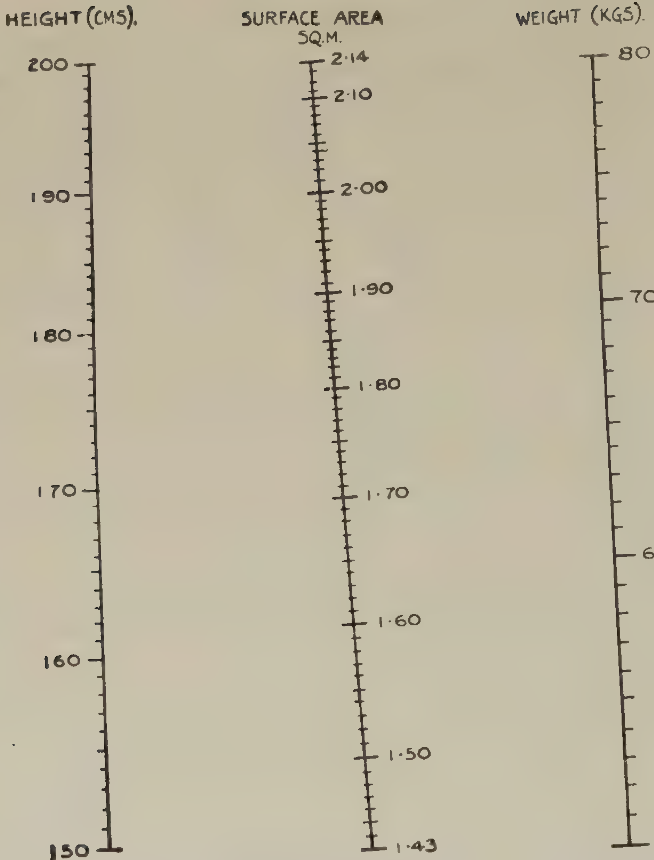


FIG. 6. 2. Alignment Chart for calculation of the area of the human body surface from the height and weight. (W. A. M. Smart.)

their biological assay. Since pure thyroxine is administered in quantities of the order of milligrams, which may increase the metabolic rate by some 200 kilocalories per day (Fig. 11. 5, p. 328), its action is clearly independent of any specific dynamic action it may have as an amino-acid. This could not amount to more than a small fraction of a calorie.

Specific Dynamic Action

Under basal conditions a certain amount of metabolism is going on at the expense of the body tissues and stores, resulting in a certain rate of energy production. If food is given to a fasting animal it is used to replace the body tissues as metabolic fuel, but in addition it causes an increase in the metabolism. For example, if sufficient protein is given to cover the calorie needs of basal metabolism, it is found that the metabolic rate rises by about 30 kcal. and the protein intake must therefore be increased, if it is to cover metabolic requirements. This effect is called the *specific dynamic action*. It is shown by fat and

carbohydrate also, but to a lesser extent. The exact cause of the specific dynamic action is uncertain, but is thought to be due to chemical changes carried out by the body on the food substances preparatory to oxidation. The specific dynamic action of protein has been most studied on account of its greater magnitude. If certain amino-acids are injected into the blood stream, they produce a specific dynamic action equal to that of the corresponding amount of protein, so that the effect cannot be due to digestion. The specific dynamic action of protein is prevented by previous removal of the liver, and this would suggest that the effect is probably concerned with deamination of the amino-acids or formation of urea rather than with stimulation of cellular metabolism generally.

The Use of Isotopes in Metabolic Studies

A new phase in the study of metabolism began with the introduction of isotopes as a means of following the chemical changes which take place in the body. Many elements can exist in more than one form and these forms are known as isotopes. The isotopes vary in atomic weight, but are chemically indistinguishable from each other, and if metabolisable compounds are prepared containing the isotopes, many experiments have shown that the animal body does not treat these in any way differently from the corresponding compounds containing the ordinary form of the element. For example, amino-acids containing nitrogen of atomic weight 15 (^{15}N) follow the same course of chemical changes as do those containing ordinary nitrogen of atomic weight 14. Two kinds of isotope are available for metabolic studies. There are first, the stable isotopes which are distinguished by the mass spectrograph and secondly, the radioactive isotopes, recognised by the radiations which they emit. Both kinds of isotope are extensively used, but in general, the radioactive ones are preferred, when possible, on account of the simpler techniques for their estimation. The use of isotopes enables certain substances, or even parts of the molecule of the substance, to be "labelled," so that they can be again recognised after passing through various chemical transformations. Sometimes two isotopes are introduced into one molecule, *e.g.* the carbon chain of an amino-acid may be labelled with the radioactive ^{14}C, while the amino group may be labelled with the stable ^{15}N or one of the hydrogen atoms replaced by deuterium. One of the great advantages of isotope studies is that chemical reactions can be investigated in whole animals under natural physiological conditions. These reactions could previously be investigated only on isolated tissues or organs, or under abnormal conditions where substances not normally metabolised by the body were used, or where inhibitors had to be used to stop metabolic processes at some particular stage.

THE METABOLIC HISTORY OF THE FOOD SUBSTANCES

In the preceding section the general methods of metabolic studies have been discussed and we can now turn to the special problems of

the various substances in the food. The energy-liberating substances are proteins, fats and carbohydrates. These are broken down in the alimentary tract, and from them are formed the substances which can be regarded as primarily available for metabolism, *i.e.* amino-acids, fatty acids and glucose. All these undergo a series of chemical transformations which result ultimately in the formation of carbon dioxide and water. (In the case of protein some other end products are also formed, *e.g.* urea.) These chemical transformations can be divided into two stages, and one of these stages—the citric acid cycle—is common to all the different types of food substances. The pre-citric acid cycle stage is quite different for different food substances ; it is therefore convenient to consider the pre-citric acid cycle stage for fats, proteins and carbohydrates, and then to consider subsequently the citric acid cycle which is common to them all (Fig. 6. 5, p. 203) The important connecting link between the two stages is a substance called coenzyme A, which therefore occupies a very special place in metabolism.

Coenzyme A. This is a complex substance containing adenosine and pantotheine (a substance related to one of the B vitamins) joined by two molecules of phosphoric acid. It exists in two forms, coenzyme A and acetyl coenzyme A, and its function can be regarded as a carrier of acetyl groups (CH_3CO—), which it is able to accept from one compound and transfer to another. Coenzyme A contains an —SH group, and since this is the part which combines with the acetyl radical, the free and combined forms of coenzyme A are often abbreviated to H—S—CoA, and $CH_3COS.CoA$. In general the final reaction in the pre-citric acid cycle stage of metabolism is the formation of acetyl-coenzyme A. This transfers the acetyl group to the citric acid cycle and coenzyme A is re-formed, to take part in the further formation of acetyl coenzyme A.

Protein Metabolism

The protein taken in the food is absorbed from the intestine after breakdown to amino-acids. Some of these amino-acids are subsequently resynthesised to protein to help to replace the worn tissues of the body, or to produce hormones or enzymes. The remaining amino-acids are used to supply energy. The replacement of the worn tissues can be carried out only by protein supplied in the food, whereas the energy supplying function of the protein can be replaced by fat or carbohydrate. Since protein is the most expensive part of the diet, and under many conditions the part most likely to be scarce, it is important to determine precisely how far it can be replaced by the other food substances, and what is the minimum protein intake. This can be discovered by studying the nitrogen balance of the body.

Nitrogen Equilibrium. When an animal is living on an adequate diet and maintaining its weight at a constant level, it takes in a certain amount of nitrogen in the protein of the food, and loses the same amount of nitrogen in the excreta. Such an animal is said to be in nitrogen equilibrium. The nitrogen in the excreta comes partly from the nitrogen of the food, and partly from the breakdown of protein

in the animal's own tissues, this being replaced again from the food. If now the animal be given a diet with ample fat and carbohydrate, but completely lacking in protein, it continues to excrete some nitrogen from its own tissues, and since this is not replaced it loses weight and ultimately dies. If after a short period of nitrogen starvation protein be added to the food in known amounts and at the same time the nitrogen loss estimated, it can readily be determined how much nitrogen must be given in the food to bring the animal back to nitrogen equilibrium. The result of such an experiment is given in Table 6. 4.

TABLE 6. 4.

Establishment of Nitrogen Equilibrium in a Dog after Starvation.

Food.	Nitrogen in Food.	Nitrogen in Excreta.	Difference.
	g.	g.	g.
Starvation	0	4·00	—4·00
100 g. Meat	4·10	5·56	—1·46
140 g. Meat	5·74	6·50	—0·76
165 g. Meat	6·77	7·22	—0·45
185 g. Meat	7·59	7·80	—0·21
200 g. Meat	8·20	8·73	—0·53
230 g. Meat	10·24	10·58	—0·34
360 g. Meat	11·99	12·05	—0·06
410 g. Meat	15·58	14·31	+1·27
360 g. Meat	13·68	13·62	+0·06
Starvation 3rd day	0	4·03	—4·03

From such an experiment two important conclusions can be drawn. First, the minimum protein requirement for the particular animal can be seen. In the example given, an intake of about 360 g. of meat is required to make the intake of protein balance the loss. But another important result appears. The amount of nitrogen which a starving animal excretes is not the amount which is needed to maintain it in nitrogen equilibrium, for it can be seen that by the time it has reached nitrogen equilibrium it is taking in and excreting about three times the amount excreted in the starvation state. The principal reason for this is the following. The protein given in the food does not supply the amino-acids in the proportion required by the body, so that food containing a considerable excess of some may have to be given to supply a sufficient quantity of others. The amount of protein needed to maintain nitrogen equilibrium depends, therefore, on the type of proteins supplied ; and the value of proteins is estimated from the amounts they contain of certain of the amino-acids, which the body cannot manufacture for itself, and which are, therefore, regarded as essential in the diet. If an animal be fed on a protein deficient in any one of these it cannot be kept in nitrogen equilibrium no matter how much of the protein is supplied. An essential amino-acid has been

defined as one which cannot be synthesised by the organism out of materials ordinarily available at a speed commensurable with the normal requirements. The essential amino-acids as determined for the rat are : lysine, valine, tryptophan, methionine, histidine, phenylalanine, leucine, isoleucine, threonine and arginine. In the human subject arginine and histidine are not essential in the adult ; while in some species glycine is essential at a certain stage of growth. From the practical aspect, the important point is that proteins can be divided into those of high biological value (first-class proteins) and low biological value (second-class proteins), the first group containing the essential amino-acids in proportions approaching those required by the body. The term " first-class protein " usually means protein of animal origin (proteins of milk, cheese, meat, eggs, etc.).

The Break-down of Protein. The amino-acids not used for synthesis of body tissues are oxidised with liberation of energy. The first stage in the process is deamination, *i.e.* the removal of the nitrogen-containing group from the rest of the molecule. If amino-acids are injected into the blood stream they are rapidly removed and can be partly recovered from various tissues. Of these the liver has been found to take up the largest quantity. It can further be shown that the amino-acids taken up by the liver gradually disappear, suggesting that they undergo some transformation. At the same time as the amino-acids disappear from the blood there is a rise in the concentration of blood urea. It is possible to keep an animal alive for some time after removal of the liver and, if in such a preparation amino-acids be injected, they are only removed slowly from the blood and there is no increased formation of urea. These experiments together with much other evidence show that the liver plays a very important part in the formation of urea from amino-acids. Urea is one of the end products of protein metabolism. It does not undergo further change and is excreted in the urine. A small fraction of the amino groups is not excreted in this form, but as ammonia which is formed largely in the kidney itself, the amount being related to the acidity of the urine.

Formation of Urea. If ammonium salts containing ^{15}N be fed to animals the isotope appears in the urea, so that formation of ammonia is probably a preliminary process in the formation of urea. The method of formation of urea from ammonia is still a matter of discussion, but it seems very likely that it depends on a cyclical series of reactions involving ornithine, citrulline and arginine. Ornithine combining with carbon dioxide and ammonia forms first citrulline and then arginine, which is hydrolysed by the enzyme arginase to form urea and ornithine.

Fate of the Non-nitrogenous Residues. The non-nitrogenous part of the amino-acid left after deamination undergoes oxidation, the ultimate products being carbon dioxide and water. In carnivorous animals much of the energy needed by the body is obtained from this source, but in man the amount obtained is relatively small, depending on the excess protein in the diet over the protein minimum. The immediate product of deamination is a keto-acid but the type of keto-

acid will vary according to the amino-acid from which it is derived. All of them are ultimately oxidised by the citric acid cycle, but before they reach that stage, the metabolism proceeds by two different routes. Some of the keto acids, *e.g.* pyruvic acid, are known to be intermediaries in carbohydrate breakdown and could thus provide a route by which the further breakdown of protein might follow the line of carbohydrate metabolism, or by which proteins could cause formation of carbohydrates. It can be shown that this does actually happen under certain conditions. In a diabetic animal, or in an animal poisoned with *phlorrhizin*, there is a great loss of sugar from the body. (Phlorrhizin is a drug which causes elimination of sugar through the kidneys without any increase in the blood sugar concentration.) In such an animal, the carbohydrate stores are rapidly depleted. If now carbohydrates be withheld from the diet but protein be given it is found that the excretion of glucose continues, and this must have been derived from protein.

The amino-acids which can give rise to glucose during metabolism are glycine, alanine, valine, serine, cystine, cysteine, ornithine, proline, hydroxyproline, aspartic acid, glutamic acid and hydroxyglutamic acid. Some amino-acids, *e.g.* leucine, phenylalanine and tyrosine, do not give rise to glucose, but to acetoacetic acid. This is a product of fat metabolism, which under certain conditions accumulates in large quantities in the blood, and leads to a condition called ketosis. (See sections on fat and carbohydrate metabolism.) For this reason, leucine, phenylalanine and tyrosine are said to be ketogenic, while those amino-acids which give rise to glucose are said to be anti-ketogenic. A few amino-acids, lysine, histidine and tryptophan form neither glucose nor acetoacetic acid.

Creatinine. This is another end product of protein metabolism. It is formed from creatine (methyl guanidine acetic acid) of which it is the anhydride. Creatine is present in the tissues in the form of creatine phosphate and, in this form, it plays a part in the chemical processes responsible for muscular contraction. Creatinine is regarded as a waste product of muscle metabolism and is always a normal constituent of the urine. Creatine is not normally present in the urine but in some cases, for reasons not understood, it may be a urinary constituent. It often appears in the urine of women, but its relation to the menstrual cycle is uncertain.

Nucleo-proteins. These are proteins found especially in the nuclei of cells, and they form a small part of the dietary protein. They consist of a protein conjugated with nucleic acid, this latter substance being a combination of a purine or pyrimidine base with phosphoric acid and a pentose. The purines present are adenine and guanine, the pyrimidines, thymine and cytosine. The nucleoproteins are broken down during digestion, and among the products are the purine and pyrimidine bases. The latter after absorption are completely oxidised. The purine bases undergo deamination and partial oxidation to uric acid. In man, this is an end product of nucleoprotein metabolism and is excreted in the

urine, but in most mammals it undergoes a further stage of oxidation to allantoin.

Exogenous and Endogenous Metabolism. The main end products of protein metabolism differ considerably in different states of nutrition. The urea output of the body is fairly closely dependent on the amount of protein in the diet. On the other hand the amount of creatinine excreted is almost constant, even during large fluctuations of dietary protein. Uric acid occupies an intermediate position as regards the relation between the amount excreted and the protein intake. The different behaviour of these substances suggests that protein metabolism can be divided into the metabolism of the protein fuel supplied by the diet, and the metabolism of that supplied by the tissues. Since the amount of creatinine does not vary with the diet it is thought to be an index of tissue metabolism, or " endogenous " metabolism, while urea is thought to be an index of " exogenous " metabolism. Experiments with isotopes have shown, however, that protein metabolism cannot be divided in this way. Animals were fed with leucine and glycine containing deuterium attached to the carbon chain and ^{15}N in the amino group of the molecule. When the excreta were collected and examined, it was found that only a small amount, about one-third, of the isotope had been eliminated from the body. An examination of the different tissues showed that the proteins of the blood, and the proteins of most of the organs, contained isotopic leucine and glycine. Since deuterium as well as ^{15}N was present in the tissue proteins it proved that not only the amino-group but the whole molecule of the amino-acid given in the food had been incorporated in the tissue proteins. The animals had not gained weight during the process, so that it was not a question of retention of amino-acids to build up extra body tissues. The only possible conclusion was that the amino-acids given had replaced some of the leucine and glycine previously present in the body proteins. This showed that there is a constant synthesis and breakdown of body protein with replacement of the amino-acid molecules by new ones derived from the dietary protein. It is thus not possible to separate the exogenous and the endogenous metabolism, as there is a dynamic equilibrium between the amino-acids in the body tissues and those in the blood stream. Further investigation showed that there was not only replacement of amino-acids in the tissue proteins, but also replacement of the nitrogen in the amino-acids. When isotopic leucine or glycine was fed, other amino-acids isolated from the tissues contained the ^{15}N. This indicated that the ^{15}N supplied in the leucine and glycine had been used for the synthesis of other amino-acids to supply new units for the tissue protein. Even when abundant quantities of some particular amino-acid are supplied in the diet synthesis of this amino-acid still occurs.

Metabolic Acidosis and Alkalosis. Most kinds of protein contain sulphur and phosphorus atoms in organic (un-ionised) combination, and some also contain chlorine atoms. When these molecules are completely oxidised, sulphate and phosphate ions (and perhaps chloride

ions) are released into the body fluids, the negative charges being derived from hydrogen atoms which are oxidised to hydrogen ions. These hydrogen ions combine with bicarbonate ions, and form carbon dioxide which is eliminated in the lungs. The net result, therefore, is the replacement of bicarbonate ions in the body fluids by sulphate and phosphate ions, so that according to the Henderson-Hasselbalch equation (Chapter 3, p. 85), if the alveolar carbon dioxide pressure were constant, the acidity of the blood would rise (the pH would fall). Actually, since the respiratory centre is activated by an increase in acidity of the blood (Chapter 4, p. 125), there would be an increase in the respiratory minute volume, the alveolar carbon dioxide pressure would fall, and the increase in acidity would be diminished. The essential point, however, is the reduction in the bicarbonate concentration of the body fluids (the " alkali reserve "), and it is this that is called *acidosis*.

For experimental purposes, this type of acidosis may be produced to almost any desired extent, by ingesting ammonium sulphate or ammonium chloride. The ammonia is converted into urea in the liver, and hydrogen ions, together with sulphate ions or chloride ions, are released into the body fluids.

Many kinds of vegetable food, on the other hand, notably fruits, contain substantial quantities of salts of organic acids. The acids undergo oxidative metabolism, with the formation of bicarbonate ions, so that the alkali reserve is increased and an *alkalosis* results. On a mixed diet, of course, the reduction in bicarbonate consequent on eating protein will be compensated by an increase on eating fruit and vegetables : but in man, the former usually preponderates, and there is a net gain of hydrogen ions, which are excreted in the urine (Chapter 9).

Lipid Metabolism

The term " lipid " is applied to a number of different classes of substance which occur in the animal body and in the diet. These are: (1) the simple triglycerides of fatty acids (the fats proper) ; (2) the sterols and their esters with fatty acids, and (3) the more complex phospholipids and cerebrosides. Of these the fats proper form the greatest bulk of the tissue and dietary lipids and most of this section will be devoted to them.

Fats. The fatty acids present in the body fat are restricted to those with even numbers of carbon atoms in the molecule. Of these all members of the series from acetic acid (2 carbon atoms) to stearic acid (18 carbon atoms) have been found, but by far the most important quantitatively are the 16 and 18 carbon atom fatty acids, palmitic ($C_{15}H_{31}COOH$) and stearic ($C_{17}H_{35}COOH$), together with the unsaturated oleic acid ($C_{17}H_{33}COOH$). These three in the form of their glycerol esters make up most of the body fats.

Unsaturated Fatty Acids. More highly unsaturated fatty acids, linoleic, linolenic and arachidonic, occur in the body and play an important though unknown part in metabolism. They cannot be synthesised in the body and,

if not supplied in the diet of rats, symptoms of deficiency appear in the form of skin disturbances.

The physical properties of the body fat depend on the varying proportions of the constituent fatty acids and are more or less characteristic for each species. This constancy of body fat is, however, dependent on dietary habit and if an animal be starved so as to reduce its fat stores and then fed with an unusual type of fat, it is quite easy to alter the nature of its body fat as regards physical and chemical properties, *e.g.* melting point, iodine number, saponification value, etc.

The Storage of Fat. There are certain parts of the body where the fat content can be greatly altered and such parts like the omentum and the subcutaneous tissues can act as fat depots. When an animal is putting on weight, fat is laid down in these stores and, when it is living on its reserves, the fat in these parts diminishes before other parts of the body are called on to contribute their share of metabolic fuel. The source of this depot fat we have just seen is the dietary fat, of which a part is used directly for oxidation and part is laid down as storage in the fat depots. If fat containing deuterium is fed to animals, it can be shown that the depots do not consist of a static deposit of storage fat, but that the fat laid down is constantly being used and replaced.

It has also been well established that the non-fat part of the diet can contribute to the fat stores. This was demonstrated by the classical experiments of Lawes and Gilbert in 1852. Young pigs were fed on a diet of barley containing very little fat, and it was found that the amount of body fat present when the animals were killed was greater than could have been obtained from the fat supplied or even the fat and protein together, thus proving that carbohydrate can be converted into fat. Whether or not fat can be derived in the body from protein is uncertain.

Interconversion of Fat and Carbohydrate. Glycerol is known to be produced in small quantities during fermentation of glucose by yeast, and it is possible that a similar production can occur in the body. The formation of fatty acids from carbohydrates takes place by addition of 2-carbon units and the substance involved in this addition is acetyl coenzyme A, which is produced from pyruvic acid. The fatty acid thus formed combines with glycerol to form neutral fat. The process of fat production from carbohydrate requires the participation of certain vitamins of the B group, and aneurin, riboflavin and pantothenic acid are all probably involved.

One important aspect of the interconversion of fat and carbohydrate in the body is the effect on the gaseous exchanges. While the respiratory quotient usually depends only on the type of food being oxidised, it will also be affected by the interconversion of fat and carbohydrate. Inspection of the formula of glucose and a typical fat shows that the latter contains fewer atoms of oxygen per carbon atom than the former and hence, when fat is being formed from carbohydrate, oxygen will be freed for use in general metabolism and so less will be taken into the body from the lungs. Since the carbon dioxide production and excretion remains unchanged the respiratory quotient

will be abnormally high. An example of this process is seen in the behaviour of those animals which hibernate. At the end of summer when they are building up large reserves of fat from carbohydrate to last over the winter, the respiratory quotient is abnormally high, and may reach a value of 1·5.

The Oxidation of Fat. The fats utilised by the body normally undergo complete oxidation with formation of carbon dioxide and water. Evidence as to how oxidation proceeds was obtained by Knoop's method of feeding to animals fatty acids containing a benzene ring attached to the carbon chain.

If benzoic acid is fed to animals it is found to be conjugated with glycine to form hippuric acid, and this is excreted in the urine.

$$C_6H_5.COOH + NH_2.CH_2.COOH = C_6H_5.CONH.CH_2.COOH$$

If phenylacetic acid is fed, the excretory product is phenaceturic acid, which again represents a conjugation with glycine.

$$C_6H_5.CH_2.COOH + NH_2.CH_2.COOH = C_6H_5.CH_2.CONH.CH_2.COOH$$

If phenyl derivatives of higher fatty acids are fed, it is found that the product of excretion is always either hippuric acid or phenaceturic acid, depending on whether there is an even or odd number of carbon atoms in the side chain. Thus hippuric acid was produced from $C_6H_5.COOH$, $C_6H_5.CH_2.CH_2.COOH$ and $C_6H_5.CH_2.CH_2.CH_2.CH_2.COOH$, while phenaceturic acid was given by $C_6H_5.CH_2.COOH$, $C_6H_5.CH_2.CH_2.CH_2.COOH.$, and $C_6H_5.CH_2.CH_2.CH_2.CH_2.CH_2COOH$. These findings suggested that during the course of oxidation of the fatty acid chain the carbon atoms are split off in pairs, or in other words, oxidation takes place at the β-carbon atom. Hence this suggested method of fat oxidation was called β-oxidation.

The details of the reactions of β-oxidation cannot be discussed here, but coenzyme A plays a very important *rôle.* In general the process can be summarised by saying that one molecule of fatty acid combines with a molecule of coenzyme A thus :—

$$R.CH_2CH_2CH_2COOH + HS.CoA \rightarrow R.CH_2CH_2CH_2COS.CoA$$

where R represents the rest of the molecule. After various intermediate stages this is oxidised to form a keto acid $R.CH_2CO.CH_2COS.CoA$.

This reacts with another molecule of coenzyme A thus :

$$R.CH_2CO.CH_2COS.CoA + HS.CoA \rightarrow R.CH_2COS.CoA + CH_3COS.CoA.$$

The result is that two C atoms are split off, with formation of acetyl coenzyme A. This process proceeds until only a four-carbon compound is left, which is $CH_3CO.CH_2COS.CoA$, *i.e.* acetoacetyl coenzyme A. The end products are thus a number of molecules of acetyl coenzyme A, together with one molecule of acetoacetyl coenzyme A, and these products are normally oxidised by the citric acid cycle.

Ketosis. In certain conditions the oxidation of fat is incomplete and certain products appear in the urine, β-hydroxybutyric acid, acetoacetic acid and acetone. These substances are called " ketone bodies," and the condition in which they appear is called ketosis. It

occurs typically in diabetes mellitus, but it also occurs in less serious disturbances such as fasting, and severe vomiting. Of these ketone bodies it is known that acetone is formed from acetoacetic acid, and that the primary substances are the four-carbon atom substances, β-hydroxybutyric acid and acetoacetic acid. Probably acetoacetic acid is formed first, but the two substances are known to be interconvertible in the liver. The appearance of ketone bodies in fat metabolism is always related to a lowered oxidation of carbohydrate by the tissues, and this may be brought about either by lack of carbohydrate as in fasting, or by inability of the tissues to oxidise carbohydrate as in diabetes. When the formation of ketone bodies is very great it may be sufficient to lower the alkali reserve of the blood and produce a condition of acidæmia (movement of pH of blood towards the acid side).

Theories of Ketone Body Formation. The metabolism of acetyl coenzyme A normally proceeds by the citric acid cycle. The acetyl coenzyme A is incorporated into the cycle by reacting with oxaloacetate to form citrate. The citrate undergoes a series of changes, one result of which is that oxaloacetate is again formed and can react with another molecule of acetyl coenzyme A. However, there are other demands on the oxaloacetate formed besides reacting with coenzyme A, and hence in order to keep the citric acid cycle going there must be a constant formation of oxaloacetate. This is in fact formed from pyruvate, which in turn is formed from glucose, and hence a constant metabolism of glucose is essential in order to provide sufficient oxaloacetate for oxidation of acetyl coenzyme A. When sufficient oxaloacetate is not available the coenzyme A formed by fat metabolism, instead of entering the citric acid cycle, forms acetoacetate. Ketosis is thus seen to result from excessive utilisation of fat relative to utilisation of carbohydrate.

Carbohydrate Metabolism

The greater part of the carbohydrate in the body consists of the glycogen stored in the tissues, principally in the liver and in the muscles, and of the glucose in the blood and tissue fluids. The total amount of carbohydrate present in a well-nourished human body, of which about one-half is present in the liver as glycogen, is equivalent to less than one day's supply at the normal rate of consumption. It must be remembered, however, that carbohydrate taken into the body can be stored as fat, and also that carbohydrate can be formed in the body from non-carbohydrate sources.

The carbohydrate absorbed from the small intestine from a normal diet consists largely of three monosaccharides—glucose, fructose and galactose. They are all convertible to glycogen in the liver, and the glycogen so formed breaks down, either by acid hydrolysis *in vitro* or under the influence of the enzyme systems present in liver, to give glucose which may therefore be considered as the form in which carbohydrate is available to the tissues.

The Equilibrium between Glycogen and Glucose. Many years ago Claude Bernard showed that the glycogen of the liver, but not that of

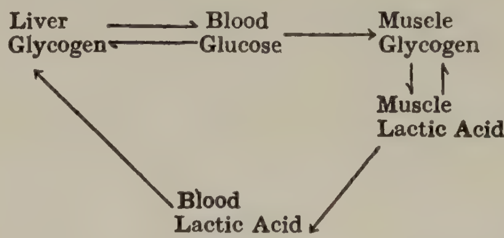
the muscles, could break down in the body to give glucose, and that, except after a recent meal containing carbohydrate, the blood leaving the liver contained more sugar than that entering the organ. The process of formation of glucose from glycogen is called glycogenolysis. Other organs of the body, and in particular the muscular tissues, appeared to be continuously absorbing glucose from the blood flowing through them. These important observations established the fact that although the concentration of the sugar in the blood falls only slightly during post-absorptive conditions or even during a long fast, this was not because the tissues ceased to absorb glucose from the blood under these conditions. The relative constancy of the blood glucose concentration was to be ascribed to the fact that the amount of glucose withdrawn from the blood by the tissues in general was counterbalanced by the amount of glucose liberated into the circulation by the liver under these conditions. The importance of this power of the liver to secrete glucose is dramatically emphasised by the fact that if the liver is removed from a dog the blood sugar concentration falls rapidly and death ensues within a few hours unless glucose is administered in large amounts. Undoubtedly some part of the glucose secreted by the liver comes from the glycogen present in it, but the stored liver glycogen is quite insufficient to supply the body's needs for very long and, as will be discussed below, glucose can be formed in the liver from non-carbohydrate sources, a process called gluconeogenesis.

After the ingestion of carbohydrate food the blood glucose concentration rises from its normal post-absorptive value of 0·08 to 0·1 per cent. to a value in the neighbourhood of 0·15 per cent. As the blood glucose concentration rises the liver secretes less and less glucose until, when the concentration reaches about 0·12 per cent., the liberation of glucose by the liver ceases altogether. At blood glucose concentrations above this, the liver begins to absorb glucose from the blood stream and the liver glycogen content begins to rise. It is clear, therefore, that the glycogen in the liver represents, in part at least, a storehouse for carbohydrate coming from the food. The glucose which is thus stored during the temporary period of plenty is liberated by the liver during the lean period of post-absorptive conditions. The ability of the liver to adjust its output of glucose to the requirements of the body is sometimes described as its homeostatic function.

Although the skeletal muscles also absorb more glucose when the blood glucose concentration rises after carbohydrate food, the glycogen which accumulates in these tissues is not reconverted to blood glucose during post-absorptive conditions. The glycogen of muscle may be oxidised or it may be converted to lactic acid, a process called glycolysis, but it never forms glucose in the body ; this contrast with events in the liver is to be ascribed to differences between the enzyme systems of liver and muscle tissues.

The process of the formation of lactic acid from glycogen in skeletal muscle is one which does not involve the addition of oxygen nor the elimination of hydrogen from the system. It is accordingly not an

oxidation process and can take place under the largely anaerobic conditions which exist in skeletal muscles during exercise. If the exercise is light, any lactic acid formed may be in part reconverted to glycogen in the muscles themselves, but if the exercise is severe most of the lactic acid may escape into the blood stream and be then converted to glycogen in the liver. Therefore, in the presence of the liver, muscle glycogen can give rise to blood glucose indirectly. These relationships will be clear from the diagram.



It should be noted that just as the formation of lactic acid from glycogen in the muscles liberates energy, the reformation of glycogen from lactic acid requires the addition of energy to the system. In the muscles any energy thus required can be provided, under aerobic conditions, by oxidation of part of the lactic acid formed or equivalent glucose or glycogen. In the liver the source of the energy required for the formation of glycogen from lactic acid is not known with certainty.

Investigations with radioactive carbon dioxide suggest that the formation of glycogen in the liver from lactic acid, and even from glucose, may involve the linking up of carbon dioxide with intermediate substances formed, so that the carbon skeleton of the glycogen which accumulates contains a significant proportion of carbon atoms originating from carbon dioxide.

Gluconeogenesis. The amount of glycogen stored in the liver is insufficient to maintain the normal level of blood sugar for more than a short time, and glucose is in fact formed by the liver from various non-carbohydrate sources, a process called gluconeogenesis. These sources are alcohols such as glycerol, many amino-acids, and a number of organic acids, *e.g.* fumaric, malic, succinic, etc., formed in metabolism. Of these the amino-acids are the most important and, hence, the dietary protein can be considered an important source of blood glucose. The process of gluconeogenesis is inhibited by insulin, and stimulated by the hormones of the anterior pituitary gland (see below).

During starvation or deprivation of carbohydrate, the utilisation of carbohydrate by the muscles and other tissues is diminished and some part of their energy requirements is met by oxidation of fats. It seems probable, however, that nervous tissue is peculiarly dependent on a constant supply of carbohydrate or similar substances and that hepatic gluconeogenesis is of particular importance in providing the nervous system with one of its requisite metabolites. It should be mentioned that the brain (and the heart also) can oxidise lactic acid

as well as glucose and thus can utilise the lactic acid formed anaerobically in skeletal muscle during vigorous exercise.

The Regulation of the Blood Sugar Concentration. The accurate control of the blood sugar concentration is of the utmost importance in the economy of the human body. If the concentration rises above about 0·18 per cent.—the renal threshold—the re-absorption of sugar from the renal tubules is incomplete and glucose is lost to the body in the urine. If, on the other hand, the concentration falls below about 0·04 per cent., the central nervous system becomes disturbed. In man there is at first extreme hunger and fatigue, with general sweating; later, delirium and profound coma are produced. In rabbits and certain other mammals, severe convulsions occur in addition just before the coma and death from respiratory failure; these are known as "*hypoglycæmic convulsions.*" If the condition is not relieved by the administration of glucose, death may ensue. Even if death does not occur as the result of a temporary lowering of the blood sugar concentration, lesions may appear in the brain which cause permanent functional damage.

The mechanism which ensures that the blood sugar concentration normally varies only within the relatively narrow range of 0·08 per cent. to 0·16 per cent., is largely hormonal in character, adrenaline, insulin and the growth hormone of the anterior pituitary being concerned.

(*a*) *Adrenaline.* When the blood sugar concentration falls below about 0·07 per cent., the excitation of nervous centres in the lower part of the brain causes sympathetic stimulation and a release of adrenaline from the adrenal glands. The adrenaline so released stimulates a breakdown of liver glycogen to glucose, the liberation of which tends to arrest the fall of blood sugar concentration. The general sympathetic stimulation may assist this process. Under the influence of adrenaline, muscle glycogen breaks down to lactic acid which may be carried to the liver for reconversion to glycogen and glucose. These several processes tend to prevent a fall in the blood sugar concentration by rapidly mobilising the glycogen stores of the body.

During muscular exercise there is a general stimulation of all the sympathico-adrenal system, so that the rate of breakdown of glycogen in the liver is accelerated on just those occasions when glucose is being used most rapidly by the muscles. As far as we know, however, adrenaline does not directly affect gluconeogenesis in the liver, but can only bring about the breakdown of glycogen which has been stored in that organ.

It was shown many years ago by Claude Bernard that puncture of the floor of the fourth ventricle in the rabbit results in hyperglycæmia and glycosuria; this is known as "*diabetic puncture.*" The hyperglycæmia is due to the excitation of fibres that run into the sympathetic system, with a consequent outpouring of adrenaline.

(*b*) *Hormones of the adrenal cortex.* The adrenal cortex plays a part in carbohydrate metabolism as is shown by the fact that if it is removed, or diseased, the blood sugar concentration falls, and there is muscular weakness, in addition to other abnormalities to be discussed in Chapter 11. Administration of the "gluco-corticoid" hormones, extracted

from the adrenal cortex, abolishes the hypoglycæmia and increases the rate of glycogen formation by the liver.

(*c*) *Insulin.* As we have seen above, when the blood sugar concentration rises as the result of the ingestion of carbohydrate food, the liver responds by first diminishing and then abolishing the liberation of sugar into the blood stream, and then proceeds to absorb part of the excess sugar in the blood stream, simultaneously increasing its glycogen content. The ability of the liver to suspend the production of sugar and inhibit gluconeogenesis, and to promote glycogen storage at the expense of the excess glucose in the blood, is dependent on the presence of insulin, the hormone secreted by the islets of Langerhans of the pancreas.

When the blood sugar concentration rises after a carbohydrate meal the excess glucose is stored as glycogen in the muscles as well as in the liver, while in the muscles the rate at which sugar is oxidised may increase under these conditions. All these processes for the storage and utilisation of glucose depend on the availability of insulin, and there is good evidence that the secretion of insulin by the pancreas is stimulated by a rise in the concentration of sugar in the blood ; the controlling action appears to be the direct effect of the excess sugar concentration on the cells of the pancreas, although it is possible that a nervous reflex is also involved. Apart from the formation of glycogen a further process for the storage of energy from excess ingested carbohydrate is the conversion of glucose to fat (glycerides) which are then stored in the fat depots. In general, therefore, insulin may be said to promote those metabolic processes which cause glucose to leave the blood stream (conversion to glycogen or fat, or promotion of carbohydrate oxidation) and to inhibit gluconeogenesis. All these processes tend to lead to a fall of the blood sugar concentration and if a large dose of insulin is given to a normal person in a post-absorptive state, fatal hypoglycæmia may develop.

(*d*) *Anterior pituitary hormones.* An animal from which the anterior pituitary gland has been removed becomes abnormally sensitive to the hypoglycæmic action of insulin. Conversely, a normal animal to which anterior pituitary extract is administered becomes highly insensitive to the action of a small dose of insulin in lowering the blood sugar concentration. These observations show that in some respects anterior-pituitary secretion and insulin act antagonistically.

It seems possible that anterior pituitary secretions exert two separate actions on carbohydrate metabolism. The first, attributed to the action of pituitary adrenocorticotrophin (A.C.T.H.), stimulates the liberation of adrenal-cortical hormones from the adrenal gland and thus promotes gluconeogenesis, with consequent accumulation of glycogen in the liver. The second action, possibly associated with the growth-promoting principle of anterior-pituitary extracts, depresses carbohydrate oxidation in the skeletal muscles with the accumulation of muscle glycogen stores. This will be discussed further in Chapter 11.

Diabetes in Man. If a normal man takes 50 g. of glucose by mouth and the blood sugar concentration is determined at intervals afterwards,

the type of curve shown in Fig. 6. 3 is obtained. The rise of the blood sugar concentration stimulates the secretion of insulin and the mechanism which we have considered in the previous section comes into play, with the result that much of the glucose is stored as glycogen ; the blood sugar concentration thus falls again without having reached so high a value that glucose is excreted by the kidneys, *i.e.* the renal threshold is not exceeded and glycosuria is absent. Such a test is

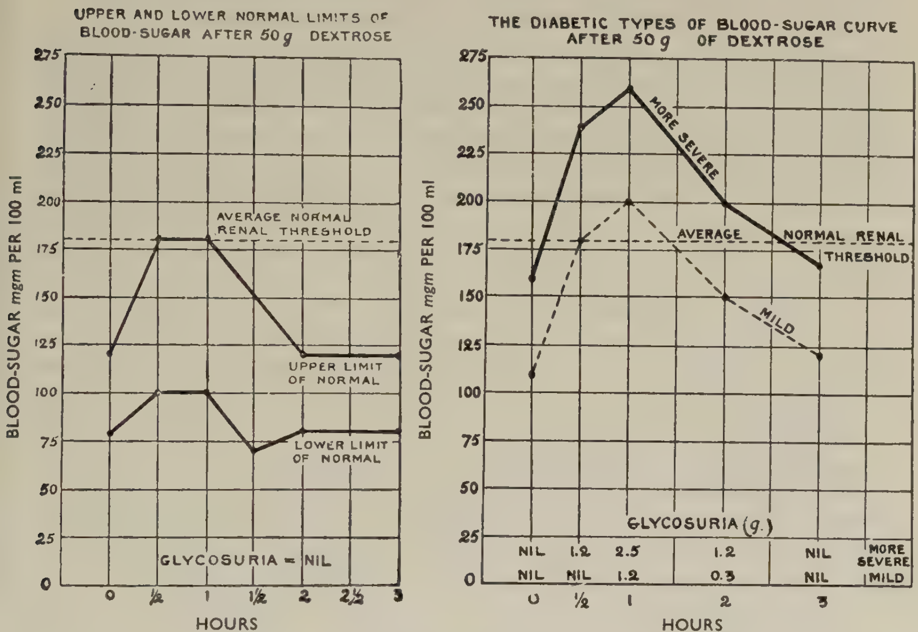


FIG. 6. 3. The effect of the ingestion of 50 g. glucose on the **Concentration of Sugar in the Blood.**

(*Left*) *Normal Subjects.* The concentration is not increased excessively and returns to the initial value within two hours. No sugar is excreted in the urine.

(*Right*) *Diabetic Subjects.* The concentration rises to a considerably higher value and takes three hours to return to the initial value. Considerable quantities of sugar are excreted in the urine.

The figures under " glycosuria " indicate the total quantity of sugar excreted in the sample of urine collected at that time. (Harrison's " Chemical Methods in Clinical Medicine.")

called a " **glucose tolerance** " test and is of much value clinically in the diagnosis of *diabetes mellitus.* Occasionally glycosuria may follow the ingestion of a large amount (150 to 200 g.) of glucose by a normal person, but this is not considered to be necessarily indicative of diabetes mellitus ; it is described as *alimentary glycosuria.*

When a glucose tolerance test is performed on a diabetic person, the blood sugar concentration may be very high initially and may rise excessively after glucose administration, failing to fall again to the

original value for many hours. The mechanism for the storage of glycogen in the liver and muscles is defective, while glucose oxidation is also subnormal ; much of the carbohydrate taken in with the food is therefore lost in the urine as glucose. The depression of carbohydrate utilisation causes the production of excess ketone bodies in the liver with the development of ketosis and ketonuria. The appearance of ketone bodies in the urine is an indication of the gravity of the disease, a slight glycosuria being of no great importance. If the ketone bodies continue to accumulate, the patient ultimately dies in a coma, if he has not already succumbed to infection, to which he becomes very sensitive. Daily subcutaneous injections of insulin, however, can relieve most of the symptoms and the patient can lead a normal life ; omission of the insulin is rapidly followed by a reversion to the diabetic condition. Excess insulin, as we have seen, leads to hypoglycæmia and serious consequences, so that doses must be carefully controlled ; immediate subcutaneous injection of glucose, or a large quantity given by mouth, quickly alleviates the symptoms should an overdose be given accidentally.

In recent years considerable advances have been made in the production of slowly-absorbed insulin preparations. These have the great advantage that fewer and larger injections can be given, and these more nearly simulate the physiological secretion of insulin by the pancreas.

The Ætiology of Diabetes Mellitus. It is extremely probable that *diabetes mellitus* is not a single syndrome with one origin ; it must consist of a number of different conditions arising from various causes but with the majority of symptoms in common. Accordingly, it is most unlikely that one causative agent can be assigned to this disease.

Experimentally a condition resembling human *diabetes mellitus* in some important respects can be induced in certain animals by surgical removal of the pancreas. The fact that degenerative lesions in the cells of the islets of Langerhans of the pancreas are sometimes found in human diabetes, together with the presence of subnormal amounts of insulin in the pancreas, suggests that pancreatic islet lesions are to be regarded as one direct cause of *diabetes mellitus*. Nevertheless, in a substantial proportion of human diabetics no obvious pancreatic islet lesions are demonstrable and, even where these exist, we still have to seek the reason for their development. Abnormally high secretory activity of the anterior pituitary gland can obviously be regarded as one possible extra-pancreatic cause of clinical *diabetes mellitus*. Experimentally, a diabetic condition can be induced in intact cats and dogs by repeated injections of anterior pituitary extract, and such treatment may cause the appearance of lesions of the pancreatic islets of such severity that they remain after the pituitary treatment ceases, so that a persistently diabetic condition is seen. After the daily injections of anterior pituitary extract have ceased the persistent diabetes is easily controlled by the administration of insulin, but during the period of pituitary treatment the diabetic condition shows a remarkable insensi-

tivity to control by insulin. It seems possible that continued overactivity of the anterior pituitary lobe may account in part for the "insulin insensitive" type of diabetes that is met with clinically, while a short period of pituitary overaction might be responsible for permanent islet lesions associated with a persistent clinical diabetes which is not abnormally unresponsive to control by insulin.

Overactivity of the secretory function of the adrenal cortex, either primary or secondary to pituitary overaction, can initiate or exacerbate clinical diabetes if adrenal steroids of the corticosterone type are produced in excessive amount, while abnormally high thyroid activity might also contribute to the development or maintenance of a diabetic condition. Thus a number of endocrine glands may be considered as possible extra-pancreatic factors concerned in the appearance of the symptoms of human *diabetes mellitus*, though dysfunction of the islets of Langerhans is, in a substantial proportion of cases, the direct, though perhaps only the secondary, cause of the condition.

The Chemistry of Carbohydrate Metabolism

The details of the chemical reactions involved in breakdown of glucose and glycogen must be sought in textbooks of biochemistry and only some general outlines will be discussed here. Like the other food substances the metabolism of glucose can be divided into two stages, the pre-citric acid cycle stage and the citric acid cycle, the connecting link being acetyl coenzyme A.

Glycolysis. The pre-citric acid cycle stage of carbohydrate meta-

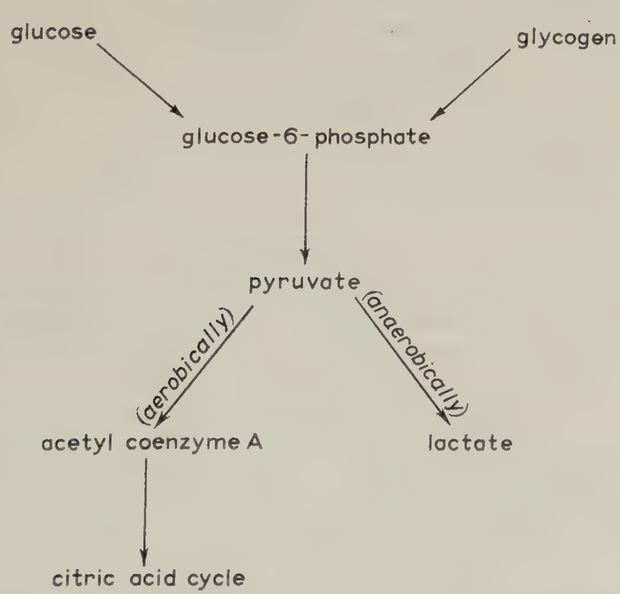


Fig. 6. 4. Carbohydrate metabolism in tissues. The intermediate stages between glycogen and glucose-6-phosphate, and between glucose-6-phosphate and pyruvate are omitted.

bolism is a series of reactions often included under the term glycolysis, although this term strictly applies to the production of lactic acid from glucose and glycogen. The relation between the oxidative pathway for glucose metabolism and the glycolytic pathway is seen in Fig. 6. 4. The reactions can start either from glucose taken in from the blood stream or from glycogen stored in the tissue. Both are converted to glucose-6-phosphate and this undergoes a series of changes which result in formation of pyruvate. In aerobic conditions the pyruvate forms acetyl coenzyme A, which goes into the citric acid cycle, while in anaerobic conditions pyruvate is reduced to lactate. This formation of lactate provides an important source of energy in muscle, when work is taking place at a faster rate than can be covered by current oxygen intake.

The Citric Acid Cycle

The citric acid cycle, or Krebs' cycle, is the final common pathway for oxidation of proteins, fats and carbohydrates. The preliminary reactions result in formation of acetyl coenzyme A, and this is the substance which enters the citric acid cycle. The details of the reactions of the cycle must be sought in textbooks of biochemistry, and only a very bare outline is given here, which is shown schematically in Fig. 6. 5. Acetyl coenzyme A (two-carbon unit) reacts with oxaloacetate (four-carbon unit) to form citrate (six-carbon unit). Citrate undergoes a series of reactions with formation and breakdown of a large number of substances and ultimate formation of oxaloacetate again. During the course of the cycle carbon dioxide appears twice as an end product, so that the six-carbon skeleton of citric acid is reduced first to a five-carbon skeleton and then to the four-carbon skeleton of oxaloacetate. The re-entry of another molecule of acetyl coenzyme A starts the cycle off again. The carbon dioxide produced by the cycle is excreted by the lungs. The other end product is hydrogen and, as seen in Fig. 6. 5, each complete operation of the cycle results in the formation of four pairs of hydrogen atoms, which are ultimately oxidised to water by the oxygen taken into the body by respiration. This final oxidation of the hydrogen is, however, in itself a complex process involving a number of different stages. The hydrogen is not able to react with molecular oxygen but is "accepted" by substances called coenzyme 1 and coenzyme 2, which are pyridine nucleotides. These substances can exist in either oxidised or reduced forms and the oxidised form can accept hydrogen and hence become reduced. The reduced forms of the coenzymes can in turn pass on their hydrogen to substances called cytochromes, which can in their turn exist in oxidised or reduced forms. The reduction of the cytochromes re-oxidises the coenzymes which are thus available to accept more hydrogen from the substances of the cycle. Finally, the cytochromes are re-oxidised by passing on their hydrogen to cytochrome oxidase and this substance is able to pass on hydrogen to molecular oxygen. The final products of the cycle are thus water and carbon dioxide.

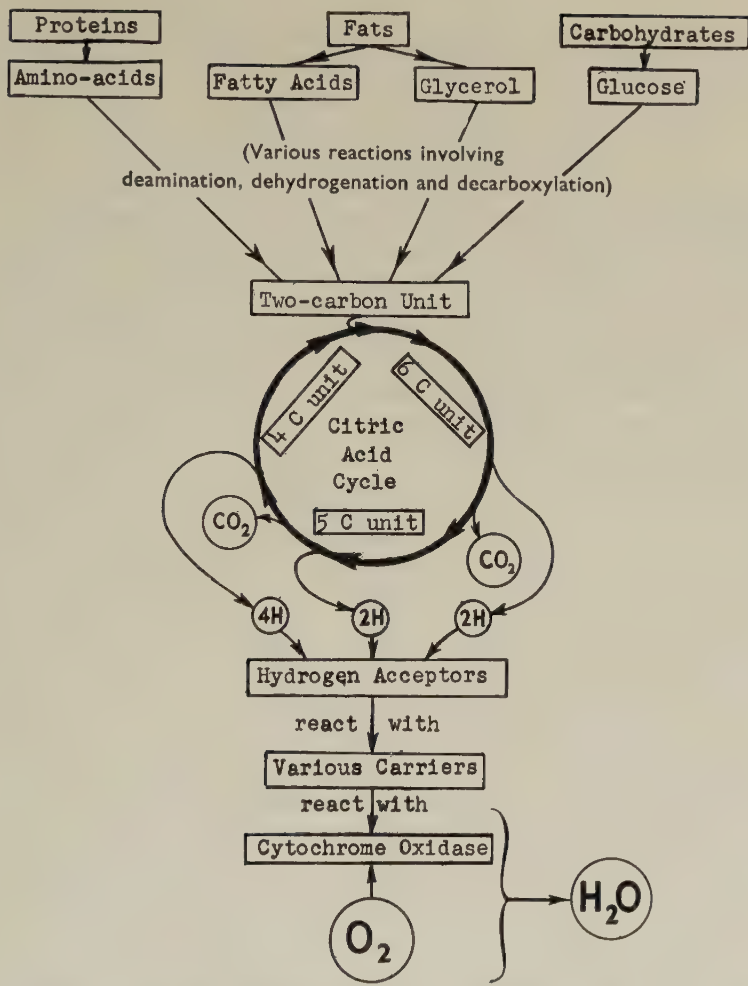


FIG. 6. 5. Schematic Representation of the Probable Course of Breakdown of the Food Substances.

The citric acid cycle is shown diagrammatically in a much abbreviated form. The various six-, five- and four-carbon compounds actually taking part form at least nine stages in the cyclical process.

Metabolism and Energy Liberation

The oxidation of food substances results in the production of water and carbon dioxide and the release of energy, which is partly utilised by the various tissues in the processes of growth, secretion, muscular contraction, etc., while the rest appears as heat. The way in which the energy becomes available is one of the fundamental problems of metabolism and it is only in recent years that a tentative answer has been formulated. It now seems fairly certain that the " useful " part

of the energy is very largely derived from one chemical reaction, the dephosphorylation of adenosine triphosphate (usually referred to as ATP). The removal of one phosphate radical from ATP results in the formation of adenosine diphosphate (ADP). The ADP formed can, in suitable conditions, combine again with a phosphate radical to form ATP. The phosphorylation process requires energy and the dephosphorylation process makes energy available. The amount of energy associated with this type of reaction is relatively large and hence, the term " active (or energy-rich) phosphate " has been used in connection with it. Since the dephosphorylation of ATP is the primary reaction which makes energy available to the tissues, the function of the other reactions of the food substances can be regarded as supplying the energy necessary for ATP synthesis ; a considerable fraction of the energy liberated by oxidative breakdown of the food substances thus becomes available. Not all oxidative reactions, even though they yield energy, can take part in ATP synthesis, and, furthermore, if the process is to be carried out efficiently, the amount of energy liberated by a particular oxidative reaction should not be too large, in relation to the amount of energy needed for ATP synthesis. This makes more understandable the apparently complicated pathway of metabolism of the food substances. The large number of reactions which occur ensures that the energy is liberated in small amounts, which are suitable for an efficient synthesis of ATP, and some of the reactions are directly " coupled " with ATP synthesis.

From the biological point of view the " efficiency " of metabolism can be measured in terms of the ATP produced ; in fact, much more ATP is generated at the citric acid stage than in the preliminary stages. Some of the ATP generated is required to maintain the metabolic reactions and, if this is subtracted from the ATP generated, the amount available for biological work is obtained, which can be called the net yield of ATP. In the case of glucose, for each molecule metabolised under aerobic conditions, the net yield in the glycolytic reactions is six ATP compared with thirty ATP in association with the citric acid cycle. Most of the ATP is generated in the oxidation of hydrogen by the pyridine nucleotide—cytochrome system. In anaerobic conditions where the pyruvate is reduced to lactate the glycolytic reactions yield only two moles of ATP.

CHAPTER 7

NUTRITION

THE natural stimuli that lead to the choice of appropriate foods are hunger, thirst and appetite, guided by the sensations of sight, taste and smell. Under the conditions of twentieth-century civilisation where food is, in general, not produced in the place where it is consumed, and where for one reason or another adequate supplies of food are likely to be periodically restricted, it is necessary to replace to some extent the operation of natural desires by carefully calculated allocation of the different food constituents. Nutrition is the science on which such attempts are based and embraces a very wide field of study. In deciding a nutritional policy for a population the following factors would have to be considered : (1) the amounts of the different food constituents required by the human body in different working and climatic conditions ; (2) the amounts of these food constituents in the different food materials used ; (3) the availability of food in terms of labour and space required for production, bulk and value of food in relation to transport, cost of food in relation to wages level, ability to import and export, etc. ; (4) the storage and preparation of food so that the highest nutritive value is maintained and (5) general propaganda and education in regard to the preparation and use of foods. Of these aspects of nutrition, only the first two belong to the province of physiology, and accordingly only these will be considered here.

The six essential classes of substance required in the diet are proteins, fats, carbohydrates, vitamins, mineral salts and water. The first problem is to know the amounts of each of these required by the body under different conditions, and the second to know the amounts of each of these in the food materials, or more accurately to know the amounts of these absorbed from the intestine. Substances originally present in the food but destroyed by cooking or not absorbed from the intestine will have no food value. In thinking of the food requirements we can divide these into foods supplying energy for bodily activity, and foods necessary for some other purpose. In the first case we are concerned with making up a certain number of calories which are required for metabolism and, in this respect, the different calorie-providing foods are within certain limits interchangeable. In the second case we have to think of specific functions of food constituents which cannot be replaced by giving other types of food. To the energy-providing foods belong proteins, fats and carbohydrates and the only other dietary constituent likely to be a source of energy is alcohol. The mineral salts, the vitamins, and water belong to the group of food substances with specific functions, but to these we must also add the proteins and fats, as these in addition to supplying energy have other *rôles*, which cannot be taken over by any other food constituent.

The Calorie Requirements

When the body is at rest it requires an energy production of about 40 kcal. per sq. metre per hour, which for an average adult amounts to about 1,700 kilocalories in twenty-four hours. When any physical activity is undertaken the energy consumption increases and the greater the physical work carried out the greater will be the need for calories. While mental work can cause a feeling of great fatigue and even hunger, it does not cause an increase in energy consumption apart from any physical movements accompanying it. Calorie requirements are, therefore, dictated by physical effort. There are two main ways in which the calorie needs can be assessed, (1) calorimetry or measurement of the metabolic rate of the body in the appropriate conditions, and (2) the dietary survey, or measurement of the amount of food consumed.

The techniques for calorimetry have already been discussed in Chapter 6 (p. 179) and, in general, indirect calorimetry, *i.e.* measurement of the oxygen consumption, is employed. Modern techniques have enabled oxygen consumption to be determined by subjects at work or play of many kinds, and as a result, fairly accurate figures are now available for calorie requirements in these conditions.

Dietary Survey. The dietary survey is an important approach to nutritional problems because it expresses calorie needs in terms of actual food used, and also gives information about wastage in preparation, storage and distribution. The survey can be carried out at different levels. There is the detailed survey on one individual, and this is the most accurate as regards actual calorie intake, provided the subject is intelligent and co-operative. Then there is the family survey, in which a trained observer attempts to assess the food intake of a normal family living under their usual conditions. In addition there is the survey at national level, where the total food production, imports, exports and consumption are assessed. These dietary surveys will give information of a rather different kind from that given by calorimetry, but this kind of information is equally important in assessing nutritional problems as a whole.

As a result of techniques of these kinds the requirements of individuals for calories can be drawn up, and in doing so it is usual to divide the day into three periods of eight hours each, *i.e.*, a period of rest in bed, a period of work, and a period of non-occupational activities, and to assess the calorie needs for each of these. This can be expressed in relation to the " reference man " or " reference woman," and these terms apply to an average individual in standard conditions. The needs of these individuals are shown in Tables 7. 1 and 7. 2 (from Food and Agriculture Organisation of the United Nations, 1957) ; corrections are then made for deviations from the standard age, weight and environmental temperature. In correcting for age in adults the following percentages of the allowances for the reference man and woman are taken : 97 per cent. for age 30–40 ; 94 per cent. for age 40–50 ; 86 per cent. for 50–60 ; 79 per cent. for 60–70 ; and 69 per

cent. for ages over 70. Children have to be considered separately, and recommended values are shown in Table 7. 3.

TABLE 7. 1

The Energy Expenditure of a "Reference Man"

(Weight, 65 kg. Age, 25 years. Mean annual environmental temperature, 10° C.)

		kcal./day
A. 8 hrs. **Working activities** : mostly standing (overall rate, 2·5 kcal./min.)		1,200
B. 8 hrs. **Non-occupational activities** :	kcal./day	
1 hr. washing, dressing, etc., at 3 kcal./min.	180	
1½ hrs. walking at about 6 km./hr. at 5·3 kcal./min.	480	
4 hrs. sitting activities at 1·54 kcal./min.	370	
1½ hrs. active recreations and/or domestic work at 5·2 kcal./min.	470	
		1,500
C. 8 hrs. **Rest in bed** at basal metabolic rate		500
	Total	3,200

TABLE 7. 2

The Energy Expenditure of a "Reference Woman"

(Weight, 55 kg. Age, 25 years. Mean annual environmental temperature, 10° C.)

		kcal./day
A. 8 hrs. **Working activities** in the home or in industry : mostly standing (overall rate, 1·83 kcal./min.)		880
B. 8 hrs. **Non-occupational activities** :	kcal./day	
1 hr. washing, dressing, etc., at 2·5 kcal./min.	150	
1 hr. walking at about 5 km./hr. at 3·6 kcal./min.	220	
5 hrs. sitting activities at 1·41 kcal./min.	420	
1 hr. active recreation and/or heavier domestic work at 3·5 kcal./min.	210	
		1,000
C. 8 hrs. **Rest in bed** at basal metabolic rate		420
	Total	2,800

TABLE 7. 3

Calorie Requirements of Children

Recommendation of Davidson, Meiklejohn & Passmore (1960)

	Age (in years)	kcal./day
Children	1–3	1,300
Children	4–6	1,700
Children	7–9	2,100
Children	10–12	2,500
Boys	13–15	3,100
Girls	13–15	2,600
Males	16–19	3,600
Females	16–19	2,400

For adjusting for different weights the following formula can be used

$$\text{For man } E = 815 + 36{\cdot}6\, W$$
$$\text{For woman } E = 580 + 31{\cdot}1\, W$$

where E is the number of kcal. and W the weight in kg. To correct for temperature the calorie requirement is increased by 3 per cent. for each 10° fall in environmental temperature from the standard, 10°, and decreased by 5 per cent. for each 10° rise. This is based on the fact

that man is better able to protect himself from the effects of cold than of heat. Occupation naturally is one of the most important factors in providing for the individual and in some heavy occupations the daily calorie requirement may amount to more than 5,000 kcal.

Calorie Content of Food. The determination of the calorie content of food by the bomb calorimeter has been discussed on p. 177, and the figures usually taken are carbohydrate and protein 4·1 kcal./g., and fat 9·2 kcal./g. Most foods eaten are in fact mixtures of these and when calorie need has to be translated in terms of actual food, reference is made to food tables which give the required values for the actual food material eaten. It is usual to find that by far the greatest part of the calorie supply in the human diet is provided by carbohydrate, as this is usually the cheapest and most plentiful source. When food is scarce, the aim is to supply sufficient protein and fat dictated by other needs (see following sections), to calculate the calories provided by these and to make up the rest of the calories with carbohydrate. A good average peace-time diet, *e.g.* as recommended by the British Medical Association Report in 1933, might make up the calories as follows :—

100 g. protein	410 kcal.
100 g. fat	920 kcal.
500 g. carbohydrate	2,050 kcal.

European diets since 1940 have certainly made up a greater proportion of the calories with carbohydrate than indicated by these figures. The general rule is that the cheaper the diet, the more carbohydrate it contains.

Specific Requirements

In the following paragraphs are discussed the specific requirements for different food substances. Figures are given in Table 7. 4 of the amounts of the various constituents which are regarded as desirable in the diet. In considering these figures it must be borne in mind that they are somewhat arbitrary and often more in the nature of a generous guess rather than an accurate knowledge of the amounts which are required. Since the effects of deprivation of certain food constituents, particularly the vitamins, may not become apparent for a long time, it is obvious that the difficulties of assessing accurately the minimum requirements of the human subject for any one food factor are very great. It is easier to be sure that a certain intake is adequate than to know what is the threshold requirement for health, and probably for this reason figures tend to be somewhat too high. They must be interpreted as something to be aimed at, rather than as a carefully determined minimum need.

Protein in the Diet. In addition to supplying energy, protein is necessary for the building up of new tissues and replacing of used ones. It has been seen in Chapter 6 (p. 187) that a minimum amount of protein must be supplied and also, that this must contain certain essential amino-acids. If even one of these is not present the body

proteins will be broken down to supply it and hence it will not be possible to keep the animal in nitrogen equilibrium. From the dietary point of view it is important to assess the nutritional value of different kinds of protein and this has been done in various ways. One way is to give a certain protein in the diet and make calculations about the fraction of the nitrogen retained by the body. This fraction is called the biological value of the protein and varies widely with different dietary proteins, *e.g.* egg albumen has a value 94 per cent., milk protein 85 per cent., gluten in wheat is as low as 40 per cent. Another method of expressing quantitatively the value of a protein is the chemical score. The amino-acid content of a protein is determined and expressed as a percentage of that of an ideal protein, *i.e.* one which contains all the essential amino-acids in the correct proportions. The lowest value found for any essential amino-acid is taken as the chemical score. There is a reasonable correlation between the biological value and the chemical score measured in these ways. A simpler approach to the problem is to classify proteins as first class (animal protein) and second class (vegetable protein) as generally, although there are exceptions, vegetable protein is of less biological value.

Recommendations have been made in various forms about dietary protein. The League of Nations (1936) suggested a minimum of 1 g./kg. body wt./day, with part of this comprising animal protein. The British Medical Association (1950) recommended that 11 per cent. of the calories of the food should come from protein (14 per cent. in the case of pregnancy, lactation, children and adolescents). The recommendations of the U.S. National Research Council are included in Table 7. 4.

Fat in the Diet. The fat of the diet is used like the carbohydrate and part of the protein to supply energy by oxidation, but in this respect it has certain advantages over the other types of food. One gram of fat on oxidation will yield more than twice as much energy as the same amount of carbohydrate or protein and furthermore the fat is taken in a concentrated form in the food, whereas, in the case of protein and carbohydrate foods, a large part of the bulk of the food is composed of water. Fat has another important *rôle* in metabolism in that it is the only form of food which can be stored in large amounts as such. This stored fat is valuable in the protection of the body against cold since it is stored partly below the skin where it forms an insulating layer. The food fat has an irreplaceable function in acting as a solvent for the fat-soluble vitamins and also for providing the body with the indispensable unsaturated fatty acids. It has been seen in Chapter 5 (p. 156) that fat inhibits the movements of the stomach and, on account of this property, fat taken in the diet prevents the onset of hunger for a longer time. It is rather striking that in spite of these important and definite *rôles* of fat in the body economy, few figures are available to suggest what are the minimum fat requirements. The usual aim is to supply 100 g. daily.

In recent years in the more highly developed countries the possibility has been widely considered of the possible harmful effects of excessive

TABLE

[1] *Recommended Dietary Allowances*
(Food and Nutrition Board,

	Kilocalories.	Protein grams.	Calcium grams.	Iron mg.
Man (70 kg.):				
Sedentary	2,500	70	0·80 (0·56)	12 (8·5)
Moderately Active	3,000			
Very Active	4,500			
Woman (56 kg.):				
Sedentary	2,100	60	0·80 (0·56)	12 (8·5)
Moderately Active	2,500			
Very Active	3,000			
Pregnancy (latter half)	2,500	85	1·5	15
Lactation	3,000	100	2·0	15
Children up to 12 years				
Under 1 year[4]	100/kg.	3 to 4/kg.	1·0	6
1–3 years[5]	1,200	40	1·0	7
4–6 years	1,600	50	1·0	8
7–9 years	2,000	60	1·0	10
10–12 years	2,500	70	1·2	12
Children over 12 years				
Girls, 13–15 years	2,800	80	1·3	15
16–20 years	2,400	75	1·0	15
Boys, 13–15 years	3,200	85	1·4	15
16–20 years	3,800	100	1·4	15

[1] Tentative goal towards which to aim in planning practical dietaries ; can be met by a good diet of natural foods. Such a diet will also provide other minerals and vitamins, the requirements for which are less well known. The restricted allowances are probably adequate for adults other than nursing or expectant mothers.

[2] Requirements may be less if provided as vitamin A ; greater if provided chiefly as the pro-vitamin, carotene.

[3] 1 mg. thiamin equals 333 I.U. ; 1 mg. ascorbic acid equals 20 I.U.

[4] Needs of infants increase from month to month. The amounts given are for approximately 6–8 months. The amounts of protein and calcium needed are less if derived from milk.

[5] Allowances are based on needs for the middle year in each group (as 2, 5, 8 etc.,) and for moderate activity.

[6] Vitamin D is undoubtedly necessary for older children and adults. When not available

7. 4

(Restricted Allowances in Brackets.)

National Research Council.)

Vitamin A[2]. I.U	Thiamin[3] (B^{1}) mg.	Riboflavin mg.	Nicotinic Acid mg.	Ascorbic Acid[3] mg.	Vitamin D I.U.
5,000	1·5 (1·1)	2·2 (1·5)	15 (10·5)	75 (52)	[6]
(3,500)	1·8 (1·3)	2·7 (1·9)	18 (13)		
	2·3 (1·6)	3·3 (2·3)	23 (16)		
5,000	1·2 (0·8)	1·8 (1·3)	12 (8)		
(3,500)	1·5 (1·1)	2·2 (1·5)	15 (10)	70 (49)	[6]
	1·8 (1·3)	2·7 (1·9)	18 (13)		
6,000	1·8	2·5	18	100	400 to 800
8,000	2·3	3·0	23	150	400 to 800
1,500	0·4	0·6	4	30	400 to 800
2,000	0·6	0·9	6	35	[6]
2,500	0·8	1·2	8	50	
3,500	1·0	1·5	10	60	
4,500	1·2	1·8	12	75	
5,000	1·4	2·0	14	80	[6]
5,000	1·2	1·8	12	80	[6]
5,000	1·6	2·4	16	90	[6]
6,000	2·0	3·0	20	100	

from sunshine, it should be provided probably up to the minimum amounts recommended for infants.

Further Recommendations. The requirement for iodine is small ; probably about 0·002 to 0·004 mg. per day for each kg. of body weight. This amounts to about 0·15 to 0·30 mg. daily for the adult, which is easily met by the regular use of iodised salt ; the use of this salt is especially important in adolescence and pregnancy.

The requirement for copper for adults is in the neighbourhood of 1·0 to 2·0 mg. per day. Infants and children require about 0·05 mg. per kg. body weight. The requirement for copper is approximately one-tenth of that for iron.

The requirement for vitamin K is usually satisfied by any good diet. Special consideration needs to be given to new-born infants. Physicians commonly give vitamin K either to the mother before delivery or to the infant immediately after birth.

fat in the diet. Fat adds greatly to the palatability of the diet, and even apart from possible specific harmful effects, fat is often a great temptation to excessive calorie intake.

Carbohydrate. As this usually forms the bulk of the human diet and is the cheapest form of calories, it is never likely to form too small a fraction of the total calorie requirement. The idea that it has no specific function is, however, erroneous. The functioning of the citric acid cycle depends on a constant supply of oxaloacetate (Chapter 6, p. 202); this is derived from dietary carbohydrate, and to a lesser extent from some amino-acids. These can be grouped together as the antiketogenic substances and, in order that ketosis should be avoided, the antiketogenic substances being metabolised should amount to at least half the ketogenic substances, which are comprised mainly of fats.

Mineral Salts. The important inorganic substances which are essential in the diet are calcium, sodium, potassium, magnesium, iron, phosphorus, iodine and chloride. In addition, smaller amounts of many other elements are required, copper, bromide, cobalt, zinc, etc. These do not liberate any energy on oxidation, but they are responsible for the maintenance of the normal function of many parts of the body. The concentration of these substances necessary in the body fluids is usually small, but on the other hand there is a constant loss of small amounts in the urine and other body secretions, and this loss must be replaced. The minerals present in highest concentration in the body fluids are sodium and chloride. In the case of most of the mineral requirements, the amounts likely to be in the diet will be more than adequate for the body's needs and there will be excretion of the excess in the urine. The substances which demand special attention are sodium chloride, iodine, calcium and iron. Sodium chloride is only likely to become deficient when there is a great loss of sweat from the body, as in conditions of working in very hot atmospheres, or in certain abnormal conditions, when there is a repeated loss of gastric juice or of intestinal secretions by persistent vomiting or diarrhœa. Iodine deficiency only occurs regionally and may cause abnormal thyroid metabolism ; this will be considered in Chapter 11. The two minerals to be considered in ordinary nutritional problems are calcium and iron.

Calcium. The recommended calcium intake can be seen from Table 7. 4. Calcium is most abundant in milk and cheese, and is contained only in very small quantities in bread and meat. It is, however, a common practice to fortify bread with increased amounts of calcium. There is a considerable loss of calcium in the intestine as part of the calcium of the food is not absorbed. For this reason the calcium required in the diet is much greater than that actually needed by the tissues. Calcium has a great diversity of functions in the body, and reference to other chapters will show its relation to heart beat, clotting of blood, clotting of milk, permeability of membranes, neuro-muscular excitability and bone formation. Pregnancy and lactation make specially heavy demands, and the dietary calcium should be specially considered in these conditions.

Iron. The chief function of iron is in connection with the hæmoglobin and the cytochrome in the tissues. In women there is a periodic loss of iron with menstruation, but in man the loss of iron from the body is extremely small. It will be recalled that disintegration of the red cells is followed by excretion of the iron-free bile pigment, while the iron is used again for hæmoglobin production (Chapter 3, p. 72). The daily amount recommended, 10 to 15 mg., certainly does not represent the amount lost from the body. It seems that most of the iron taken in the food is not absorbed, but a certain surplus is necessary in order that a small fraction should be available. Deficiency of iron is associated with anæmia.

The Vitamins

It is well known that animals cannot be maintained in good health on diets which will supply the necessary calories together with protein, fat and mineral requirements, if certain accessory food factors, the vitamins, are absent from the diet. There are, at present, a great many different substances which have been recognised as vitamins. The usual conception of a vitamin is an organic substance which is necessary for health, including growth in young animals, and which does not act by supplying energy. Most of the vitamins have a known composition and chemical formula and many of them can be synthesised. Several of the vitamins, especially of the B group, are known to take part in the oxidative processes of the body, either as coenzymes or carriers (aneurin, nicotinic acid, riboflavin).

The vitamins can be classified according to their solubility in fats or in water. The following list shows the different members of each of these groups.

- Fat soluble :—
  - Vitamin A
  - Vitamin D_2 (calciferol)
  - Vitamin D_3
  - Vitamin E (alpha-tocopherol)
  - Vitamin K
- Water soluble :—
  - Vitamin B group
    - Thiamine (aneurin)
    - Riboflavin
    - Pantothenic acid
    - Nicotinic acid amide (P.P. factor)
    - Pyridoxine
    - B_{12} (cobalamin)
    - Biotin
    - Choline
    - Folic acid
  - Vitamin C (ascorbic acid)
  - Vitamin P (citrin)

The human dietary requirement of the more important vitamins is given in Table 7. 4.

Vitamin A. Deficiency of vitamin A in the diet leads to cessation of growth, loss of weight and decreased resistance to infection. There is keratinisation of the epithelium in the eye, respiratory tract and genito-urinary tract. In the human subject there is xerophthalmia and night-blindness. The vitamin is related to β-carotene, and represents half the molecule of this with addition of an alcoholic group. β-carotene can be regarded as a forerunner of the vitamin and can replace it in the diet, if large quantities are given. Neither β-carotene nor vitamin A can be synthesised in the animal body, but they are ingested with green plants and are found in the fatty tissues, especially the liver. There is a chemical relation between vitamin A and rhodopsin (visual purple) which probably explains the connection between vitamin A and night-blindness.

Sources : Butter, egg yolk, carrots, spinach, and particularly fish liver oils.

Thiamine (Aneurin). This is the anti-neuritic part of the B complex, and its absence leads to polyneuritis in the pigeon and rat, and to beri-beri in man. Lesser degrees of deficiency cause fatigue, loss of appetite, dyspnœa on exertion, and neuritis. In the pigeon and rat, bradycardia (slowing of the heart) is characteristic. Experimentally and clinically, symptoms of deficiency occur on a diet composed mainly of polished rice, as the vitamin is present in the outer part of the grain. The pyrophosphoric acid ester of thiamine is co-carboxylase, which forms part of the enzyme system for metabolising pyruvate, and in deficiency of the vitamin the pyruvate concentration of the blood is raised. Thiamine is also essential for the conversion of carbohydrate into fat.

Sources : Yeast, wheat germ, pulses, meat.

Riboflavin. Deficiency of riboflavin in the diet causes disturbances in the mouth and tongue, in the cornea and in the skin. In combination with a protein and phosphoric acid it forms flavo-proteins, enzymes concerned with tissue respiration.

Sources : Yeast, liver, meat, eggs.

Nicotinic acid (P.P. factor). This is the pellagra-preventing factor of the B group. Pellagra is a disease characterised by diarrhœa and skin disturbances in the human subject. In dogs, deficiency of the vitamin causes black-tongue. Either nicotonic acid or its amide will prevent these disturbances. Chemically nicotinic acid amide forms part of the molecule of the phospho-pyridine nucleotides, known as Coenzymes I and II, or as DPN and TPN, substances which are known to take part in the oxidative processes in the body.

Sources : Yeast, meat, fish, wheat flour, liver.

Pyridoxine. Deficiency of this substance has been found to produce a disturbance in young rats, which resembles pellagra, but which cannot be cured with nicotinic acid : it can be obtained from rice bran. Pyridoxine forms part of the enzyme systems necessary for protein metabolism.

Pantothenic acid. When rats are fed on diets deficient in the B complex with added aneurin, nicotinic acid, riboflavin and pyridoxine they develop a greying of the skin, which can be prevented by addition of pantothenic acid to the diet. Deficiency of pantothenic acid also leads to a pellagra-like dermatitis in chickens. The function of pantothenic acid in human nutrition is not known, but it is thought to be essential. Pantothenic acid enters into the constitution of Coenzyme A which plays a major *rôle* in metabolism (Chapter 6, p. 186).

Vitamin B_{12}. This has already been referred to in Chapter 3 in connection with the life history of the red blood corpuscles. It is the anti-pernicious anæmia factor, present in liver, and is identical with the "extrinsic factor." The molecule contains about 4 per cent. of cobalt. Its relation to the other factors concerned in the maturation of the erythrocytes is not yet fully known.

Other vitamins of the B group. Other substances which have been found to be effective in replacing deficiencies in experimental diets are choline, folic acid, biotin, *p*-aminobenzoic acid and inositol. Choline is one of the constituents of the phospholipid, lecithin. If it is not present in the diet, changes occur in the liver which are chiefly characterised by excessive fat deposition. The addition of choline to the diet prevents this, and choline is thus said to exert a "lipotropic action". Little is known about the human requirements of the other substances.

Vitamin C (ascorbic acid, antiscorbutic vitamin). It has been known for several hundred years that scurvy could be prevented by including fresh fruits in the diet and it was found in 1932 that the anti-scorbutic substance in fresh fruits was ascorbic acid. In the animal body ascorbic acid is found in the adrenal cortex. On oxidation it readily forms dehydro-ascorbic acid, and can be easily reformed from this by reduction, but the biological significance of this is unknown. Vitamin C deficiency is associated with disturbance in the formation of the enamel in the teeth, and with the process of calcification of bone.

Sources: Fresh fruits and vegetables, especially the citrus fruits, oranges and lemons.

Vitamin D (antirachitic vitamin). Absence of vitamin D in the diet gives rise to the characteristic appearance of rickets, a disease of children associated with softening of the bones and hence giving rise to abnormal shapes of the parts of the skeleton which have to bear the weight of the body. Rickets is associated with an abnormal metabolism of calcium and phosphorus, and it can be prevented by giving vitamin D. The antirachitic property is possessed by a number of substances, but of these the most important are called vitamin D_2 and vitamin D_3. Vitamin D_3 (cholecalciferol) is the naturally occurring vitamin, while vitamin D_2 (ergocalciferol) is produced from ultra-violet irradiation of ergosterol. Vitamin D_3 can also be produced by ultra-violet irradiation, the precursor in this case being 7-dehydrocholesterol. This substance is present in the skin and the beneficial effect of sunlight in preventing rickets is due to formation of the vitamin. The mode of action of these substances in preventing rickets is not known. Vitamin D is the

vitamin which is most likely to be inadequate in the diet and hence the widespread habit of giving to children cod liver oil or other vitamin D source.

Sources : Butter, cream, eggs, but especially fish liver oils.

Vitamin E (alpha-tocopherol). Deficiency of vitamin E in rats gives rise to failure of the reproductive organs. In male rats there is deterioration of the testis and in the female, death of the fœtus. In other animals deficiency may be accompanied by muscular disturbances. Little is known definitely about the requirements of the human subject, and the value of treatment with vitamin E for prevention of abortion is still a matter of some dispute.

Sources : Cereals, especially oats and wheat, liver and eggs.

Vitamin K (anti-hæmorrhagic vitamin). Deficiency of vitamin K in the diet of chickens produces a hæmorrhagic disturbance associated with a lowering of the amount of prothrombin in the blood (Chapter 21). In the human subject a vitamin K deficiency can be produced when there is an absence of bile in the intestine, *e.g.* in case of a biliary fistula. The vitamin does not have any effect on such hæmorrhagic diseases as hæmophilia or purpura and, as far as is known, is only related to formation of prothrombin.

Sources : It occurs more abundantly in plants than in animals. Cabbage, spinach, cauliflower are good sources, while milk is very poor.

Alcohol

Alcohol is capable of oxidation by the animal body, but it is usually taken not so much for the purpose of providing energy as for the effect it produces on the higher centres of the central nervous system. The relaxation of rigid self-control and discrimination which it produces in suitable doses is found by many people to increase the enjoyment of congenial company, and its widespread use lends interest to a consideration of its metabolism by the tissues of the body.

Alcohol differs from other energy-providing foods in that it can be absorbed from the stomach, although the rate of gastric absorption is much less than that from the intestine. After absorption most of it is metabolised, the remainder being excreted either in the urine or in the expired air. When taken in small doses so that the concentration in the blood does not rise above a certain level, most of the alcohol is oxidised without any accompanying pharmacological action.

The oxidation of 1 g. of alcohol in the body yields 7 kilocalories of energy. Alcohol metabolism proceeds at a practically constant rate for any one individual, varying usually from 6 to 10 g. of absolute alcohol per hour. The food value of alcohol is, however, limited in spite of the fact that it is quickly absorbed and requires no digestion, partly because there is no storage mechanism and partly because of the inconvenient effects of alcohol on the central nervous system.

If alcohol is taken in larger doses it acts on various parts of the body, chiefly on the central nervous system. The effect is mainly dependent on the concentration in the blood, although it is also influenced by

the rate at which the concentration is attained. Hence the effects of a large dose of alcohol will be reduced if it is taken with food and particularly with fatty foods which delay the emptying time of the stomach. Thus cocktails before a meal are more potent than liqueurs containing a similar quantity of alcohol after the meal and the practice of starting a meal with hors d'œuvres with a high content of fat and oils is conducive to the retention of a discriminating palate throughout the course of that meal. The relation between blood concentration and pharmacological action is of considerable medico-legal interest, as the symptoms of alcoholism can be roughly related to the concentration of alcohol in the blood at the particular time (Fig. 7. 1).

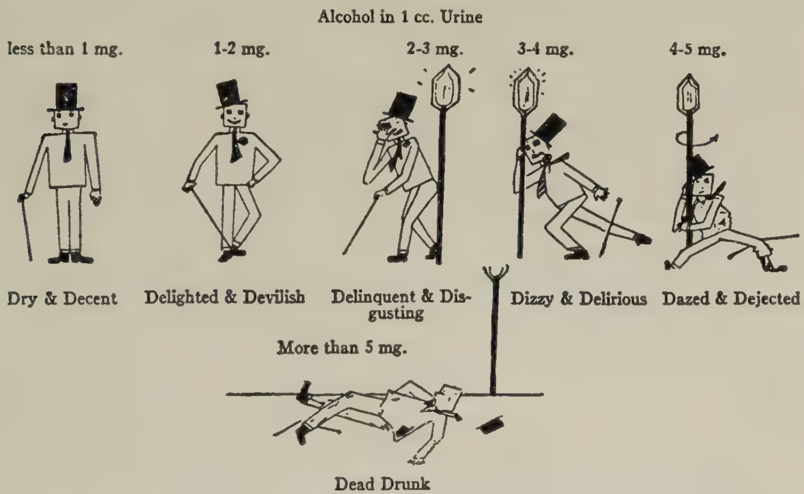


FIG. 7. 1. The relation of the **Degree of Intoxication** to the **Concentration of Alcohol** in the urine.

The concentration of alcohol in the urine is approximately the same as that in the blood and tissues, except during absorption from the alimentary canal, when the concentrations are changing rapidly. (From Emil Boger Emerson's " Alcohol and Man," by permission of The Macmillan Co.)

Alcohol has a definite diuretic action. If 50 g. ethyl alcohol are taken in 250 ml. of water it is followed by an output of urine of 600 to 1,000 ml. in two to three hours. The mechanism of the diuresis may be depression of the hypothalamic centre with consequent decrease in the secretion of the anti-diuretic hormone of the pituitary gland (Chapters 9 and 11). Alcohol is neither concentrated nor diluted by the kidney and the concentration in the urine may therefore be used as a rough measure of the concentration in the blood plasma.

Hunger and Appetite

The sensations of hunger and appetite are ill-defined feelings of " emptiness " in the one case and pleasant anticipation of food in the other. Although usually related, they are more or less distinct sensations. The hunger feeling is referred to the epigastrium and may be

accompanied by contractions of the stomach. Attempts have been made to relate hunger to some measurable change in the blood, *e.g.* to the blood sugar concentration. That there is no simple relationship is shown by the observations that injection of insulin produces a fall in blood sugar which is accompanied by a sensation of hunger, while on the other hand hunger is a common symptom of diabetes where the blood sugar is raised. Appetite is still more difficult to relate to any physiological or biochemical basis. It is more susceptible than hunger to other influences such as emotion, habit or artificial stimulation by attractively prepared food. Under conditions of modern life other factors besides hunger and appetite take part in the selection of food—in young people training and example, in adults advertising and propaganda. Fashions, fads and the cultivation of a discriminating palate also play a part in the choice of diet.

It is now recognised that the hypothalamus (see Chapter 15) plays an important part in the regulation of food intake. Experimental lesions in the hypothalamus in rats may produce a voracious appetite, and such animals rapidly increase in weight ; other lesions may abolish appetite. Electrical stimulation of the hypothalamus (in unanæsthetised goats) has led to excessive feeding. The most probable basis for hypothalamic control of appetite is that the hypothalamus responds in some way to the total amount of fat in the body, with increase in fat content depressing food intake. How this mechanism functions is quite unknown.

How far the sense of taste is a reliable nutritional guide is a problem of considerable interest. A good deal of experimental work in this field has been done on animals suffering from deprivation of some body constituent, and given the opportunity to select a diet from a large choice of substances with a view to testing their ability to make good the deficiency. Removal of the adrenal gland causes a fatal loss of sodium from the body, and it was found that adrenalectomised rats chose sodium salts out of a number of substances available and by this process outlived a control group of adrenalectomised rats to which sodium was not given. Sodium-deficient sheep, also, choose to drink salt solutions rather than water. In another set of experiments, rats from which the parathyroid glands had been removed increased their intake of calcium lactate, and thereby prevented the fall in blood calcium which, accompanied by tetany, usually supervenes in parathyroidectomised animals.

One of the important principles of animal physiology is the maintenance of the " internal environment " of the body cells and many reflex processes contribute to this end. It has been suggested that the ability to select diets suitable to physiological need is an example of the behaviour regulators acting towards the common goal of stabilising the composition of the body fluids, so essential for the continued activity of the cells (physiological " homeostasis ").

Water

Water is an essential constituent of a man's diet : he can live longer without food than without water. The amount that he must take each day, as such, by drinking, depends on the amount which is lost from the body and on the amount which is gained in other less obvious ways.

A continuous loss of water is unavoidable, for the following reasons. (1) The non-volatile end-products of metabolism—chiefly those of nitrogenous metabolism—and any excess of salts taken in with the food, are eliminated in the urine dissolved in water. Normally, a man loses about 2 litres of water a day in the urine : but this may be reduced to less than 1 litre a day when the loss of water has been greater than the intake, as in hot, dry environments ; or increased many-fold if excess water has been taken.

(2) About 0·1 litre/day is lost in the fæces. This small quantity is the residue of the much larger quantity of water which enters the alimentary canal—2·1 litres with the food and drink (see below) and possibly up to 9 litres with the digestive juices (Chapter 5, p. 146)—most of it being absorbed : it may vary considerably according to the condition of the large intestine.

(3) About 0·4 litre/day is lost by evaporation in the lungs, the expired air being saturated with water vapour at 37° C., while the inspired air is only partially saturated at a much lower temperature : this loss will be greater when the air is very dry, and will be many times as great when the ventilation of the lungs is increased, as for example during muscular activity.

(4) About 0·9 litre/day is lost by evaporation from the skin by "insensible perspiration," the skin usually being warmer than the air and not completely "water-tight." Some loss of water is necessary in order to get rid of the heat produced by the combustion of the food. Normally, in temperate climates about one-quarter to one-third of the total heat produced is lost by evaporation, but in hot surroundings nearly all of it is lost in this way, and there is manifest sweating (Chapter 20). During muscular exercise, when the metabolic rate is greatly increased, some 10 or 15 litres of water per day may be lost from the skin.

A gain of water, on the other hand, is also unavoidable, even if none is drunk. (1) The molecules of all food substances contain hydrogen atoms, and these on oxidation give rise to water molecules : an ordinary diet, yielding, say, 3,000 kilocalories a day, produces about 0·36 litre of water in this way.

(2) Water is concealed in apparently "dry" food substances : the ordinary diet provides in this way some 1 to 2 litres of water a day. (This includes the water added to the food when preparing it for the table.)

These losses and gains of water are summarised in Table 7. 6.

Some calculations of the gain and loss of water in the metabolism of various food substances, in ordinary conditions in temperate climates are given in Table 7. 5.

In constructing this Table, it has been assumed, to begin with, that the quantity of each foodstuff eaten is such as to provide 100 kilocalories of energy in its metabolism. We then know, from its average composition, how much water is " concealed " within it and how much water is formed when it is oxidised. It is then assumed that one-quarter of the total quantity of heat produced—*i.e.* 25 kcal.—is lost by evaporating water, so that, knowing its latent heat, we know the amount of water evaporated. Lastly, in order to estimate the minimum quantity of water necessarily lost in excreting the end-products of metabolism and the mineral constituents of the food, the following assumptions are made : (*a*) the urine contains 4 per cent. of urea, provided that the salt concentration is less than 0·5 per cent. ; (*b*) the urine may contain up to 2 per cent. of salts ; (*c*) as the salt concentration rises, the urea concentration falls progressively and is only 2 per cent. when the salt concentration is also 2 per cent. These assumptions are based on experimental observations, but are somewhat arbitrary, and the values calculated can be only approximate.

TABLE 7. 5

Quantity of water lost or gained when 100 *kilocalories are released in the complete metabolism of certain food substances.*

Food Substance	Gain of Water (grams)		Minimum Loss of Water (grams)		Net Loss (grams)
	Preformed	By oxidation	In dissipating heat (1/4 of total)	In excreting end-products	
Protein	0	10·5	43	212	245
Starch	0	13·5	43	0	30
Fat	0	11·5	43	0	32
Meat (lean) (average)	52	11	43	150	130
Fish (lean) (*e.g.* cod)	100	10·5	43	230	143
Fish (fatty) (*e.g.* herring)	33	11	43	85	84
Eggs (whole)	48	12	43	80	63
Bread (white)	18	13	43	40	52
Milk (cow's)	130	12	43	70	−29
Bananas	67	14	43	37	−1
Apples (fresh)	225	13·5	43	65	−130

The figures emphasise the need for fluid water in the diet, as it is seen that of the substances listed only fresh apples and, in the conditions assumed, cow's milk, provide any extra water over and above that which must be lost from the body as a result of taking the food. In drier and warmer climates (indoors as well as outdoors) than those of the British Isles, such as those of the U.S.A., a larger quantity of water is lost by evaporation, up to 60 grams for each 100 kilocalories of heat produced : cow's milk may then fail to provide any appreciable excess of water. In human milk, however, there is a smaller concentration of protein and a greater concentration of carbohydrates ; except perhaps in tropical conditions, human milk will always provide an excess of water, even to an infant whose kidneys cannot produce so concentrated a urine as can those of an adult.

Water Balance. As first stressed by Claude Bernard in 1879, the concentration of the fluid surrounding the living cells of all vertebrates and most invertebrates (the " internal environment ") remains constant, within quite narrow limits. Water can move freely into and out of the cells, so that the concentration of the fluid inside them is also constant. The concentration of the body fluids is determined, of course, by the amount of dissolved substances (chiefly salts) present in them, as well as by the amount of water. The two fluids, inside and outside the cells, contain different kinds of solute, however, as will be brought out in Chapter 8 : although salt balance and water balance are interrelated, it is important that the amount of water in the body should not fluctuate very greatly in spite of the unavoidable losses and gains already discussed. In fact, as brought out in Table 7. 6, the gains are adjusted so as to equal the losses.

TABLE 7. 6

Water Balance. *Representative values for 24 hours in a temperate climate.*

Loss (*litre*)		*Gain* (*litre*)	
Urine	1·5	Food—	
Fæces	0·1	preformed	1·0
Lungs	0·4	oxidative	1·4
Skin	0·9	Drink	1·5
Total	2·9	Total	2·9

A man weighing 70 kg. contains about 14 litres of water outside the cells (in the internal environment), so that about one-fifth of this is ordinary lost and replaced each day. In exceptional conditions, the daily " turnover " may amount to almost the whole of this extracellular volume. An infant, of 7 kg. weight, will have an extracellular fluid volume of about 1·6 litre, but the daily intake and loss of fluid will be about 0·7 litre even in normal circumstances. One-half of the extracellular fluid is thus " turned over " each day. It is clear that the intake must be accurately controlled if water balance is to be preserved, and that in the infant the margin of reserve is much smaller than in the adult ; any variation in the water intake and water loss will cause proportionately greater disturbances in the volume of its body fluids. The relative values of the volume of the extracellular fluid and of the volumes of water gained and lost per day are summarised in Fig. 7. 2.

Water balance must ordinarily be maintained by drinking an appropriate amount of water and the need to do this is indicated by the sensation of thirst. There is no special sensation for indicating that too much water has been drunk, but any excess " spills over " rapidly through the kidneys. This " water diuresis " is brought about by an entirely involuntary reflex initiated by " osmoreceptors " sensitive to the concentration of the body fluids, and controlling the release of the

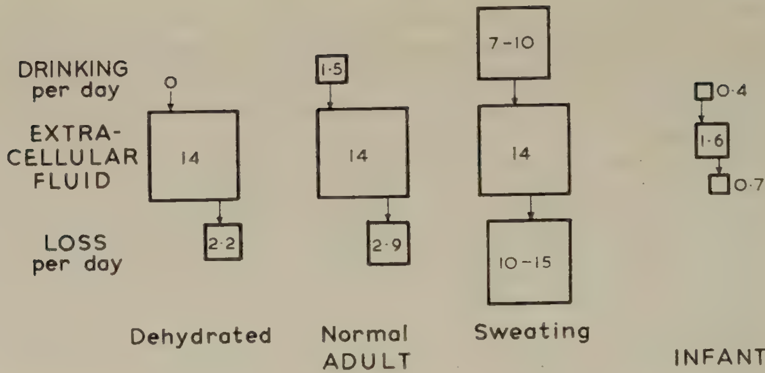


FIG. 7. 2. Diagram illustrating **Water Balance**

The volume of water taken in by drinking is the difference between the total volume lost and the volume derived from the food. The dehydrated adult would not be in water balance, the net loss of water being about 0·8 litre/day, even though the urine is as concentrated as possible.

" antidiuretic hormone " by the neurohypophysis (see Chapters 9 and 11).

Disturbances of the Water Balance. Loss of water from the body may be affected by several kinds of disorder, notably in the activities of certain ductless glands, as discussed more fully in Chapter 11. Deficiency in the secretion of the neurohypophysis (post-pituitary gland) leads to an excessive excretion of dilute urine (*diabetes insipidus*) and thus to dehydration of the body. Deficiency in the secretion of the adrenal cortex leads to an excessive excretion of sodium chloride and secondarily to a loss of water. Depletion of the body fluids may also result from : (1) loss of fluid from the alimentary tract by persistent vomiting or diarrhœa ; and (2) metabolic disturbances such as *diabetes mellitus* (Chapter 6, p. 199), when there is an unavoidable loss of water and salts accompanying the glucose and ketone bodies excreted by the kidneys. Disorder of the kidneys or of the heart, on the other hand, may so reduce the rate of elimination of water and salts that the volume of the body fluids becomes excessive, as described in Chapter 8.

In abnormal external conditions, the water balance may become disturbed even in normal men. Sustained and profuse sweating, for example, involves large losses of both water and salts both of which must be replaced. If only water is taken, the body fluids become diluted, leading to water diuresis and loss of much of the water. Correspondingly, many people find that thirst is quenched most effectively after copious sweating if they take enough salt along with the water. If a normal man drinks too much water, and for some reason the excess is not adequately eliminated, " water intoxication " may follow, as in " miners' cramp " (Chapter 20, p. 600). On the other hand, if he drinks only sea water, in which salts are more concentrated than they are in the most concentrated urine that can be formed, the body fluids

inevitably become excessively concentrated ; as they may do in people surviving from shipwreck after the fresh water is exhausted.

Most of the experimental investigations on the control of water and salt balance must be performed on conscious animals or human subjects, because most general anæsthetics put the controlling systems out of action. The experimental procedures that can be used are correspondingly limited, and interpretation of the results made more difficult.

Thirst

It is generally agreed that water intake is largely governed by thirst, indeed it is difficult to define or measure thirst except in terms of the volume of water required to assuage it. Nevertheless, there is no good agreement about the physiological basis of the sensation of thirst except to the extent that several factors must be involved.

Dehydration of the body tissues becomes manifest most conspicuously as dryness of the mouth and throat, and it has been thought that sensations from this region give rise to a drinking reflex. There is now substantial doubt about this simple explanation of thirst partly because dehydrated animals with œsophageal fistulæ, so that no water reaches the stomach, drink more than they need to rehydrate the body ; it seems that some sensation of fullness in the stomach is concerned in inhibiting drinking. Indeed, dilatation of a balloon in the stomach will put a stop to the drinking of such experimental animals. Moreover, a fairly high proportion of normal people claim to feel thirsty when the mouths are not dry, and quite a few fail to feel thirsty when their mouths are dried by drugs such as hyoscine, employed to prevent travel sickness but which unfortunately also inhibit salivary secretion.

An altogether different approach to the problem of thirst has been the search for a *drinking centre* in the brain, activated by dehydration of its cells which therefore act as osmoreceptors. (These appear to be different from those which are concerned in the onset of water diuresis.) Injection into a confined region of the hypothalamus of a goat of a few microlitres of a solution of sodium chloride more concentrated than the body fluids, induces excessive drinking ; while similar injection of a solution of the same concentration as the body fluids, or less concentrated, does not. Moreover, electrical stimulation of the appropriate region induces drinking which begins 20 to 30 seconds after the stimulus starts and lasts until a few seconds after it stops. Over-hydration up to 40 per cent. has been produced in goats in this way. Furthermore, destruction of the appropriate region by electrocoagulation has abolished drinking in dehydrated dogs and rats. Clearly, therefore, a region in the hypothalamus is concerned with drinking behaviour and may incorporate the essential osmoreceptors which signal dehydration.

If such a centre were accepted as the primary control of drinking, the usual association of drinking with sensation of dryness in the mouth and throat could well be interpreted as a conditioned reflex (Chapter 14, p. 447), as could the inhibition of drinking by the sensation of fullness in the stomach. On the other hand, these and other receptors

which respond to tissue dehydration and the volume of water drunk may give rise to primary reflexes starting and stopping drinking independently of the central receptors associated with the "centre" in the brain.

Profound hæmorrhage makes most people thirsty and this has been attributed to reduction of the volume of fluid in the body, exciting some unidentified volume receptors. After hæmorrhage, or reduction of the fluid volume by other means, rats have been found to drink more : but no consistent change in drinking pattern has followed substantial hæmorrhage in horses or dogs, nor in human donors who have yielded from 5 per cent. to 10 per cent. of their blood volume to the blood banks. Presumably, therefore, the traditional belief that thirst follows profound hæmorrhage from wounds or disease such as gastric ulcer, is to be explained largely in terms of some other shock-like effect of the injury, and not only as a consequence of the reduction in blood volume.

CHAPTER 8

BODY FLUIDS

THE greater part of the bulk and weight of any animal consists of watery solutions ; the structural part, which gives the animal its solidity, has less than one-half the weight of the fluid part. It is in the body fluids that most of the chemical reactions involved in metabolism take place, and it is by means of diffusion in these fluids, and by their movement as a whole from place to place, that the metabolites reach and enter the active cells and the products of the metabolic reactions leave and are carried away. Changes in the volume and composition of these fluids may thus affect the activities of most of the organs and tissues of the body, and it is becoming increasingly apparent that the study of these changes in health and disease is an important branch of physiology.

The composition of this fluid matrix is not uniform throughout, and variations in the concentrations of different kinds of dissolved substances are found in different parts of the body. It is useful to think of the animal body as divided into distinct fluid compartments which are more or less separated by various barriers, but at the same time are in equilibrium with each other—or if not in true equilibrium, in a " steady state," maintained by appropriate metabolic reactions. The whole body fluid may be conveniently divided into two major compartments : (1) the intracellular fluid enclosed by the cell walls, and (2)

TABLE 8. 1.
Body Fluids
(*in an average man, weight* 70 *kg.*)

	Kilogram	Per cent. of body weight
Total Body Fluids	49	70
(1) Intracellular Fluid	35	50
(2) Extracellular Fluid	14	20
(*a*) Interstitial Fluid	11·2	16
(*b*) Blood Plasma	2·8	4

the extracellular fluid between and around the cells ; this latter can be again divided into (*a*) the interstitial, or tissue, fluid and (*b*) the blood plasma. The interstitial fluid, which lies between the cell walls and the walls of the blood vessels, forms the " internal environment " for the cells, and provides the connection between the intracellular fluid, where the metabolic reactions occur, and the plasma. The plasma, by virtue of its circulation, distributes the local changes in the composition of the interstitial fluid which are produced by metabolic activity, and enables those produced in the muscles, liver, etc., to be offset by those produced in the lungs, kidneys and alimentary tract.

Each of the different fluid compartments has its characteristic volume and composition. Their relative sizes are given in Table 8. 1, and their compositions are discussed in a later section.

Methods of Estimating the Fluid Volumes

A known quantity of some suitable substance is added to the fluid whose volume is to be estimated and its concentration measured : the quantity added, in grams for example, divided by the concentration, in grams per litre, is the volume of the fluid in which it is dissolved, in litres. The principle of the method is thus simple, but discovering the " suitable substance " is not so simple.

The substance chosen must be harmless and without pharmacological effects which would change the amount or distribution of body fluids. In practice, it must be injected into the blood stream, allowed to become distributed throughout the body, and its concentration estimated in a sample of venous blood, usually taken from an arm vein. Corrections must be made for any production or destruction of the substance in the body, as for example by metabolic processes, and for the quantity eliminated *e.g.* in the urine, during the interval between injection and collection of blood for analysis. It is obvious that these corrections should be small if possible. Many of the most accurate methods depend on the use of radioactive isotopes ; special facilities for handling and estimating radioactive substances are needed and proper precautions must be taken against radiation hazards.

(1) **The Volume of the Blood.** The " reference substance " must be retained within the blood vessels for a period at least long enough to allow it to become uniformly distributed. This is ensured by using a substance which is attached either to the plasma proteins, the volume of the plasma being measured ; or to the red blood cells, the total volume of these cells, or that of the whole blood, being measured. The volumes of the plasma, of the red cells and of the whole blood are related to each other by the hæmatocrit value (volume of red cells as per cent. of volume of whole blood—Chapter 3, p. 70), so that any one may be calculated from any other. The volume of the whole blood, however, is best measured by adding together the volume of the plasma and the volume of the red cells. Three methods have been used most frequently.

(*a*) *Plasma Volume.* A known quantity of the non-toxic dye Evans Blue T 1824, which becomes bound to the serum albumin, or a known quantity of serum albumin which has been iodinated with radioactive iodine (^{131}I), is injected intravenously. The blood sample is centrifuged and the concentration of dye, or of " labelled " protein, in the plasma is estimated.

(*b*) *Red Cell Volume.* A quantity of red blood cells which have been " labelled " by means of radioactive isotopes of iron, phosphorus or chromium (Chapter 3, p. 71), and of known radioactivity, is injected. The radioactivity of a known volume of red cells, packed by centrifuging at high speed, is measured.

(*c*) *Total Blood Volume.* A known amount of carbon monoxide mixed with oxygen is re-breathed until all the carbon monoxide is absorbed and combined with hæmoglobin. The amount of CO-hæmoglobin in unit volume of blood is estimated by means of the reversion spectroscope (Chapter 3, p. 82)

(2) **The Volume of the Extracellular Fluid.** The substance injected must now be able to pass through the walls of the blood capillaries, but unable to pass through the cell membranes. Many different substances have been used, but none satisfies the criteria perfectly and all have some disadvantages. Reliable values seem to have been obtained by the use of inulin, mannitol, sucrose (in man only), sodium sulphate (containing the radioactive isotope of sulphur for ease of analysis), sodium thiosulphate and sodium thiocyanate. The volume which is measured is that of all the water in which the reference substance is dissolved ; the volume of the whole extracellular space is larger than this, since there are microscopic and ultra-microscopic structures into which the substance does not penetrate (see p. 229).

(3) **The Volume of the Intracellular Fluid.** This can only be measured as the difference between the total volume of the body fluids and the volume of the extracellular fluid (apart from the volume of the red blood cells, as already described). The measurements are thus subject to considerable uncertainty. The intracellular fluid contains substances of large molecular weight and its total volume is greater than that of the intracellular water.

(4) **The Total Volume of the Body Fluids.** Fundamentally, the most perfect method of measuring the volume of water in an animal is to weigh it, kill it, dry the body in an oven, and weigh it again. But if the animal, or man, is not to be destroyed by the method of estimation, a reference substance must be used which penetrates into all the water of the body—*i.e.* into all three compartments. Very few substances have yet been discovered which are at all suitable. Water containing the isotopes of hydrogen (deuterium, or radioactive tritium) is theoretically the best, but the analyses require elaborate apparatus, and tritium is not easily obtained. Urea may be used, although the corrections for metabolism and excretion are rather large and certain other organic compounds (the drug antipyrin, for example) have been found which are better.

The Composition of the Body Fluids

Water is the largest single component of the body fluids. Its presence in the proper quantity is important for two reasons : first, to maintain the total volume of the body fluids, and the volumes of the various compartments, and secondly, to maintain the concentrations of the solutes in some, at least, of these fluids. A great many different substances are dissolved in the water, and in general, the different compartments have different compositions. Some substances are present in all the compartments, but few of them have the same concentration in all.

The processes by which the water content of the body is maintained, and a " water balance " achieved, have been discussed in Chapter 7. We shall be concerned here with the nature of the dissolved substances in the various compartments and with the movements of these, and of the water, from one compartment to another.

Extracellular Fluid. Apart from proteins, the interstitial fluid has very nearly the same composition as the blood plasma and, since enough plasma for chemical estimation can easily be collected, its composition is usually taken as equivalent to that of the extracellular fluid. The concentrations of the diffusible electrolytes (sodium, potassium, chloride and bicarbonate), however, are not quite the same in the plasma as in the interstitial fluid, as will be discussed later.

The peritoneal, pleural and pericardial fluids, although separable anatomically from the bulk of the interstitial fluid, all have much the same composition and may be regarded as parts of it. Their particular function is largely one of lubrication, enabling the intestines, lungs and heart to move freely in their respective enclosures. Lubrication of the articular surfaces of the bones is performed by the **synovial fluid** ; this differs from the other fluids in containing mucin, which helps in the lubricating action.

The **cerebro-spinal fluid** and the **intra-ocular fluids** are usually regarded as part of the extracellular fluid, but differ quantitatively in composition from the general interstitial fluid. They will, therefore, be considered separately, the cerebro-spinal fluid at the end of this chapter, and the intra-ocular fluids as essential parts of the eye in Chapter 18. Fluid is also present in the inner ear and labyrinth ; that which lies between the bony wall and the membranous labyrinth—the **perilymph**—has the composition of the interstitial fluid, but that which lies within the membranous labyrinth—the **endolymph**—has a very different composition. This will be discussed in Chapter 19, in connection with the physiology of hearing.

In spite of the fact that there is no difficulty in obtaining quite large quantities of human blood, a full and exact description of its chemical composition cannot be given. It contains many substances whose identity is known, but in too small a concentration for accurate estimation ; and other substances whose presence can be inferred only from their actions, many of them not having been isolated or chemically identified, for example some hormones and enzymes.

Intracellular Fluid. The cells which make up the various organs and tissues are intimately surrounded by connective tissue and interstitial fluid ; an organ or tissue which has been dissected out, however cleanly, includes both intracellular and extracellular fluids. In order to discover the composition of the intracellular fluid from the gross composition of the whole organ or tissue, it is necessary, therefore, to discover the fraction of the whole organ or tissue which consists of extracellular fluid. This can be done by the use of a reference substance which cannot penetrate into the cells, as described already (p. 227) : but since complete accuracy is impossible, we know even less about the exact composition of the intracellular fluid than about that of the extracellular fluid. Moderately complete analyses are available for mammalian striated muscle and much less complete analyses for some other mammalian organs and tissues. The red blood corpuscles do not present this complication, since they can be obtained practically free from plasma by centrifuging the blood, but in some respects their composition is not typical of intracellular fluid in general.

Proteins. The blood plasma contains some 7 to 8 grams of protein in solution in each 100 ml. This protein can be split into many different fractions, but for most purposes it is sufficient to consider it as being composed of two parts, serum globulin and serum albumin, accounting for 2 g. and 5 g. per 100 ml. plasma respectively, in addition to which there is about 0·3 g. to 0·4 g. of fibrinogen. The functional significance of these proteins will be considered in later chapters. The interstitial space, on the other hand, contains only small quantities of protein in solution ; but, particularly in the connective tissues, there are substantial quantities of proteins of various kinds, chiefly collagen and elastin, in the form of fine fibres visible under the microscope, which provide rigidity and elasticity to the whole structure. These are surrounded by the " ground substance " which contains muco-polysaccharides (polymerised amino-sugars), notably hyaluronic acid, as well as proteins. The " intercellular cement " lies between, and holds together, the individual cells of an organ or tissue, and has the same (or a very similar) composition. The presence of extracellular material which is not in solution is particularly obvious, of course, in such structures as hairs, finger- and toe-nails, tendons, teeth, cartilage and bone.

Proteins in quite considerable concentration also occur within the cells, both in solution and as microscopically visible structures ; the distinction, however, is bound to be somewhat vague, since the electron microscope reveals as structures, fibrils for example, what is invisible in the light microscope. Some kinds of protein, notably the nucleo-proteins and those which form many kinds of enzyme, are found in all kinds of cell. The nucleo-proteins in many kinds of cell appear ordinarily to be in solution, since they are invisible : but during mitosis, when the cell divides and becomes two cells, some of the nucleo-proteins become visible as chromatin threads and chromosomes. Other kinds of protein are peculiar to the particular kind of cell considered, and

are essential to the function of those cells. The red blood cells, for example, contain hæmoglobin in large concentration but apparently in solution ; this is essential to their function of transporting oxygen and carbon dioxide, as discussed in Chapter 3. Muscle cells contain

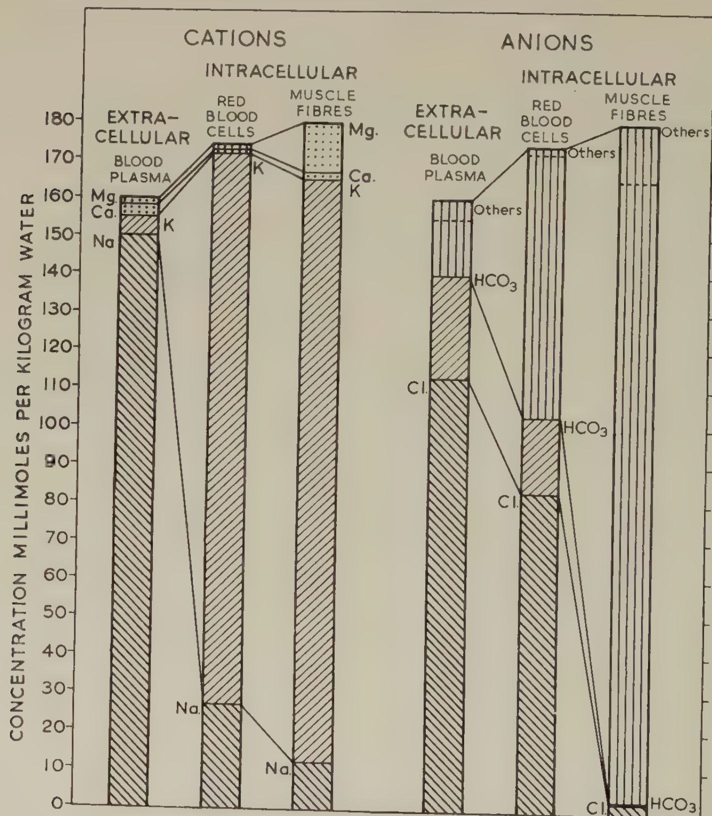


FIG. 8. 1. The Concentration of Electrolytes in the Extracellular and Intracellular Fluids of Man.

The values plotted may be taken as representative, but there are appreciable variations between one individual and another, and in any one individual according to circumstances.

The concentrations are given in millimoles per kilogram of water ; those of the " other anions " are deduced from the necessity for the total anion concentration to be electrically equivalent to the total cation concentration. The intracellular fluid appears to be more concentrated than the extracellular fluid ; but actually the two fluids are osmotically equivalent, as some of the intracellular constituents are not osmotically active.

special kinds of protein, in relatively small concentration, but in the form of fibrils and filaments ; these are essential for the shortening and development of tension characteristic of muscle cells, as is discussed in Chapter 16.

The concentration of the proteins, in the intracellular fluid and in the blood plasma, in terms of grams per unit volume of fluid, is quite

large ; the weight (and volume) of the water present is thus correspondingly less than that of the whole fluid. In 100 ml. of blood plasma, there are 90 to 93 grams of water, and in 100 ml. of red blood cells or muscle fibres, there are 70 to 80 grams of water.

Electrolytes. All the body fluids contain salts, and electrolytes, in solution, but their concentrations in the intracellular fluid are very different from those in the extracellular fluid and it is necessary to consider separately the cations and the anions which together make up these salts. The concentrations of the chief cations and anions in the blood plasma and in the intracellular fluid of red blood cells and muscle fibres are plotted diagrammatically in Fig. 8. 1.

When considering the chemical properties and osmotic and electrical equilibria of a solution, the concentrations should be expressed in terms of gram-molecules rather than in grams, and per kilogram of water, rather than per litre of solution ; the substances are dissolved in the water and not in the proteins or other colloidal constituents which may occupy an appreciable fraction of the whole volume of the fluid. A millimolar solution of a substance, say sodium chloride, contains 1/1,000 of its molecular weight in grams (*e.g.* 58·5 mg.) in 1 kilogram (or approximately 1 litre) of water. Concentrations are commonly estimated, and expressed, nevertheless as mg./100 ml. of solution.

Cations. The greatest difference between the extracellular fluid and the intracellular fluid is that the former contains chiefly sodium, the molar ratio Na/K being about 30, whereas the latter contains chiefly potassium, the molar ratio Na/K being about 0·18 in the red blood cells, and about 0·08 in the muscle fibres. This very large excess of potassium is characteristic of muscle and nerve fibres, but other kinds of cell may be more similar in this respect to the red blood cells.

In some kinds of animal other than man, the red blood cells do not contain an excess of potassium, the Na/K ratio in the cells of the cat and the dog, for example, being about 15.

The only other cations present in significant, though relatively small, concentration are calcium and magnesium. The calcium concentration of the intracellular fluid is somewhat less than that of the extracellular fluid and the magnesium concentration is definitely greater ; both are present within the cells largely in combination with organic substances, and not in the free state as ions.

Anions. In the extracellular fluid the most important anions are chloride and bicarbonate, the other anions which are present, sulphate, phosphate and lactate, together making up only some 2 per cent. of the total concentration. Within the cells, however, the combined concentration of chloride and bicarbonate is far too small to be equivalent to that of the cations. It is necessary to suppose, therefore, that other kinds of anion are present in substantial concentration. In red blood cells, these are provided almost entirely by hæmoglobin ; in muscle fibres, they consist mainly of hexose phosphates, creatine phosphate, adenosine triphosphate and the substance carnosine (β-alanyl histidine),

some also being provided by proteins. The concentration of protein ions will depend on the acidity (or pH) of the fluid, since this will affect the strength of the protein as an acid ; there will then be inverse changes in the bicarbonate concentration. This effect is of considerable importance in the regulation of the acidity of the blood and the carriage of carbon dioxide, as discussed in Chapter 3.

Other Substances. The extracellular fluid contains, in easily analysable concentration, glucose, urea, and " lipids " such as lecithin and cholesterol. Urea is the only normal constituent of the body fluids which is present in the same concentration (about 0·03 g./100 ml. or 5·5 mM) in both intracellular and extracellular fluids ; its concentration depends on the amount of protein taken in the diet, so that there may be quite large differences between different individuals. The glucose concentration in the blood (about 0·1 g./100 ml. or 5·5 mM) is maintained relatively constant, as discussed in Chapter 6. It is converted in the intracellular fluid into hexose phosphates and glycogen, for example, but glucose may be present in some kinds of cell.

The Interchange of Water and Dissolved Substances between the Various Compartments

It is a fundamental law of physical chemistry, based on universal experience of the properties of matter and energy, that the components of a solution tend to move from regions where they are in high concentration, to regions where they are in low concentration. They will, in fact, move in this direction unless they are prevented from doing so by some obstruction, or are driven in the opposite direction by some other force. This force may be electrical ; it may be mechanical, such as a hydrostatic pressure ; or it may be chemical, such as would be produced by interactions with other substances present in the solution or in contact with it. This general law applies to the water as well as to the substances dissolved in it. The concentration of the water becomes smaller as the concentration of the dissolved substances becomes larger, since, in a given volume, water molecules are replaced by solute molecules. But when we refer to the " concentration " of a solution, we ordinarily mean the *solute* concentration : water will thus move from a *less* concentrated solution (in this sense) to a *more* concentrated solution. The concentrations, however, must be expressed in terms of " osmolarity," that is, of the sum of the (molar) concentrations of *all* the solutes present, each ion of an electrolyte being considered as a separate substance. Unless the solution is very dilute, we must then multiply by the appropriate value of the " osmotic coefficient," which may be found in Tables of Physical and Chemical constants. For mammalian body fluids, the value is about 0·9. The osmolar concentration of any solution may be measured directly in terms of its freezing point, or vapour pressure.

The fact, as indicated in Fig. 8. 1, that the fluids in the different compartments have different compositions, shows that there must be obstructions, or barriers of some kind at their junctions. Between the

intracellular and interstitial fluids there are the cell membranes, and between the interstitial fluid and the plasma there are the walls of the capillary blood vessels. It is the properties of these barriers which determine the nature of the interchanges which occur between the fluids in the different compartments.

The Donnan Membrane Equilibrium. Suppose we have two solutions of sodium chloride, and add to one of them the sodium salt of a protein or of any of the organic anions found within living cells, as mentioned above. The two solutions are now put one on each side of a boundary, or membrane, through which sodium and chloride ions can penetrate but the organic anions cannot. The concentration of sodium ions is greater on one side of the membrane than on the other, but their natural tendency to diffuse in the direction of the concentration gradient is immediately opposed by an electrical force set up by the indiffusible anions which are left behind ; this pulls them in again. Chloride ions, on the other hand, are driven out, so as to create a concentration gradient, by the same electrical force ; a force which pulls in positively charged ions will push out negatively charged ions. The freely diffusible ions thus become distributed unevenly between the two solutions, and an electrical potential difference is developed across the membrane. In these conditions, all freely diffusible ions which may be present (sodium, potassium, chloride and bicarbonate, for example) will move from one solution to the other until, for each kind of ion, the concentration in one solution is related to that in the other by a general equation defining what is known as the " Donnan Membrane Equilibrium." Using plasma (*pl*) and interstitial fluid (*int*) separated by the capillary wall as concrete examples, the equation is :

$$\frac{[Na^+]pl}{[Na^+]int} = \frac{[K^+]pl}{[K^+]int} = r \; ; \quad \text{and} \quad \frac{[Cl^-]pl}{[Cl^-]int} = \frac{[HCO_3^-]pl}{[HCO_3^-]int} = \frac{1}{r}$$

the square brackets indicating concentrations. The electrical potential difference between the two fluids is given by :

$$\begin{aligned} E = RT/F . \ln . r &= 61{\cdot}5 \log . r \text{ at } 37^\circ \text{ C. (millivolts)} \\ &= 57{\cdot}2 \log . r \text{ at } 15^\circ \text{ C. (millivolts)} \end{aligned}$$

The value of the " distribution ratio," denoted by r, increases if the concentration of the indiffusible (*e.g.* protein) ions is increased, or if the concentration of the diffusible (e.g. chloride) ions is decreased. In the plasma, the indiffusible ions are anions, and r is greater than 1, the concentrations of sodium and potassium are greater in the plasma than in the interstitial fluid (by about 5 per cent.) and the concentrations of chloride and bicarbonate are smaller ; the plasma is electrically negative to the interstitial fluid by about 1 millivolt.

The equation, as just given, is an approximation only, since instead of concentrations we should, strictly, use " activities." By direct analysis of plasma and of interstitial fluid, or of a simple solution in equilibrium with plasma at a membrane impermeable to proteins (an " ultra-filtrate " or " dialysate " of plasma), it is found that the value of r for sodium ions is 1·06, and that from chloride ions is 1·04. The apparent discrepancy is due to the fact

that the activity coefficients of sodium and chloride ions are slightly smaller in the plasma than in a protein-free fluid.

Interchanges across the Cell Membranes. We are still very far from understanding adequately the nature and properties of the cell membranes. Certain facts, however, which are of importance in the present connection are well-established. Water can penetrate relatively freely into and out of the cells ; certain substances in solution, notably oxygen and carbon dioxide, can penetrate nearly as freely ; other substances, such as urea and certain organic compounds (known as " non-polar " or " lipoid soluble " compounds) whose molecular weight is not too large, and small univalent ions, can penetrate, but much less freely.

The non-polar substances are those which are soluble in ether, benzene and other liquid hydrocarbons, in which the " lipoid " materials, such as long-chain fatty-acids, cholesterol and lecithin, are also soluble. It is partly because of the relative ease with which non-polar substances penetrate, that the cell membrane is thought to be composed chiefly of lipoid material. For this reason, also, non-polar substances are used for determining the total volume of water in the body.

From the compositions of the intracellular and extracellular fluid given in Fig. 8. 1, we see that the concentration gradients of potassium ions and of chloride ions are in the direction which is to be expected if there is a Donnan equilibrium set up by the presence of indiffusible anions within the cells. There is no reason to suppose that these ions (and bicarbonate ions) cannot penetrate the cell membranes. Indeed, the use of radioactive isotopes and the existence of the " chloride shift " in the red blood cells (as discussed in Chapter 3) show that they do penetrate. There is also an electrical potential difference across the cell membranes of the sign to be expected (intracellular fluid negative to extracellular fluid) and of about the right size (10 to 100 millivolts). But the concentration gradient of sodium ions is in the opposite direction from that of potassium ions and it might seem, at first sight, that the cell membranes must be totally impermeable to sodium ions. But the use of isotopes has shown that this is not so, and it is impossible to avoid the conclusion that sodium ions are being expelled from the interior of the cell by some active process, the " sodium pump," which is only kept going by a continuous supply of energy from metabolic reactions. In effect, the cell membranes behave as if they were impermeable to sodium ions, since any that enter are immediately expelled.

A " sodium pump " which expels positively charged ions will generate an electrical potential difference between the intracellular fluid and the extracellular fluid, and this will draw out chloride ions as well. The concentration ratios of all the ions which can diffuse freely will still be defined by the membrane equilibrium equation, as given above ; and the value of the distribution ratio r will be related to the electrical potential difference across the membrane whether this is due to the presence of indiffusible ions or to the action of the " sodium

pump." But the outward movement of sodium ions through the membranes of most kinds of cell is at least partly " coupled " to the inward movement of potassium ions, one potassium ion going in for each sodium ion " pumped " out. (Such a " pump " will not generate an electrical potential difference.) Owing to this restraint on the movement of potassium ions, the concentration ratio (intracellular fluid)/(extracellular fluid) is not precisely the inverse of that of chloride ions, nor precisely that to be expected from the electrical potential difference across the cell membrane. The distribution of potassium ions, as well as that of sodium ions, between the intracellular fluid and the extracellular fluid is not that to be expected if there were a Donnan equilibrium between the fluids, and the left-hand part of the membrane equilibrium equation does not apply. In red blood cells, the discrepancy is large, for both sodium ions and potassium ions, as may be seen in Fig. 8. 1 ; but in muscle and nerve cells the discrepancy is small for potassium ions, though large for sodium ions. This will be discussed further in Chapter 12.

The cell membranes thus allow many substances in solution to pass through them. But if there is a change in the total osmolar concentration of either the intracellular fluid or of the extracellular fluid, water is found to pass from one to the other much more rapidly than any of the substances in solution. The cells swell or shrink until the two fluids are once more in equilibrium with one another. Since the cells are not rigid, and can swell or shrink quite freely, changes in the concentration of the extracellular fluid are accompanied by equal changes in the concentration of the intracellular fluid, and *vice versa* ; the consequent shifts of water from one to the other may, even in the normal living animal, be of quite considerable magnitude. For example, the activity of almost any organ is accompanied by the production of metabolites which are mainly of smaller molecular weight than the parent substances from which they are derived ; the osmolar concentration of the active cells rises, therefore, and the cells swell. Muscular exercise, in particular, causes a rise in the volume of the intracellular fluid which can be observed, in man, by the methods of measurement already described.

Owing to the maintenance of osmotic equilibrium between the intracellular and extracellular fluids by a shift of water from one to the other, it is important that the substances used for measuring the volume of the extracellular fluid should be such that they can be analysed accurately even in low concentration. If the substance added to the extracellular fluid increases the osmolar concentration significantly, enough water will be drawn out from the intracellular fluid to produce a significant error in the estimated volume of the extracellular fluid. Indeed, by suitable calculation, it is possible to deduce the *total* volume of water in the body from the concentration of an added substance which does not enter the cells at all.

Isotonic Solutions. A solution is defined as being " isotonic " if, when placed in contact with living cells, no water passes into or out of the cells, " hypertonic " if it draws water out of the cells, and " hypotonic " if water goes into the cells. If two solutions are separated by a

membrane permeable only to water, and impermeable to all the substances in solution, water will move from the solution with the smaller total molecular (osmolar) concentration to the solution with the larger osmolar concentration. But if the membrane is permeable to some of the substances in solution, these substances will move from one solution to the other until their concentrations are the same in both, and they will then have no effect on the movement of water. In calculating whether a solution is isotonic or not, the concentrations of all those substances to which the cell membranes are permeable must be left out of consideration. (This may not be strictly true if the substances

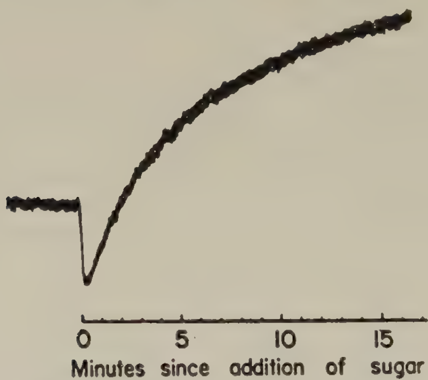


Fig. 8. 2. Swelling and Shrinking of Human Red Blood Cells.

Changes in the average volume of the cells were recorded directly in terms of the amount of light transmitted through a very dilute suspension in a saline solution. At time 0, 2 ml. of a solution of sorbose was added to 10 ml. of the suspension, the final concentration of sorbose being 0·3 M. The total osmolarity of the suspending solution was thus approximately doubled. The optical transmittance decreased suddenly, indicating an almost instantaneous shrinkage of the cells : movement of water between the cells and the outside solution is very rapid.

The sorbose then penetrated slowly into the cells, raised the concentration of the intracellular fluid, water was drawn in, and the cells swelled. The final optical transmittance is greater than the initial, since (*a*) the whole suspension has been diluted to 5/6 (the effect of the initial shrinkage is partly masked by this) ; and (*b*) the suspending solution has been made hypotonic, since water was added as well as sorbose and the cells have swelled accordingly. (Lefevre and Davies.)

are electrolytes, owing to the effect of the Donnan equilibrium, but it is very nearly true.) For example, the red blood corpuscles are impermeable to sodium chloride, but permeable to urea, oxygen and carbon dioxide, so that the addition, of, say, urea to an isotonic solution of sodium chloride will not make this solution hypertonic with respect to the red blood corpuscles. Again, the ordinary collodion membrane, like the glomerular membrane in the kidney, is permeable to all crystalloidal substances, so that only the concentrations of the colloidal ones (such as proteins) need be considered. All these membranes, however, are more permeable to water than to any solute, so that transient osmotic effects may be observed even with solutions that are, in

the long run, isotonic with respect to these membranes. In physiology, we often use the word "isotonic," without further specification, to mean "isotonic with respect to the red blood corpuscles." For mammalian cells, a solution containing 0·90 to 0·95 grams of sodium chloride in 100 ml. of water is "isotonic."

The movements of water and dissolved substances into and out of the intracellular fluid are very conveniently studied by using suspensions of red blood corpuscles. Changes in their volume can be measured by the use of the hæmatocrit (Chapter 3, p. 70) or, as in Fig. 8. 2, in terms of the amount of light transmitted by the suspension. Records from an experiment which illustrates the points just discussed are given in Fig. 8. 2. If a solution of the hexose sugar sorbose, of such a concentration as to be apparently hypertonic, is added to a suspension of red blood cells in an isotonic saline solution, there is initially a very rapid withdrawal of water from the cells, which therefore shrink. The sorbose then slowly penetrates into the cells, drawing water with it. Finally the cells are more swollen than they were initially, since the saline solution has been diluted by the water added with the sorbose, while the sorbose, having the same concentration inside and outside the cells, is osmotically inactive.

An isotonic solution may, alternatively, be defined as a solution which has the same osmotic pressure as the intracellular fluid, a hypertonic solution as one with a greater osmotic pressure, and a hypotonic solution as one with a smaller osmotic pressure. But it must be remembered that the osmotic pressure exerted by a solution depends on the nature of the membrane at which it is developed as well as on the concentration of the solution. The "tonicity" of a solution must therefore be defined in terms of the osmotic pressure exerted at the particular membranes considered—*e.g.* the membranes of certain kinds of cell, the glomerular membranes, etc.—and not necessarily in terms of the "ideal" osmotic pressure exerted at a membrane permeable only to water.

Hæmolysis. If the cell membranes are destroyed, or sufficiently damaged, the cell contents pass into the surrounding solution. The red blood cells, again, are very useful for studying the conditions in which this occurs: the hæmoglobin which they contain has a strong red colour and is thus easily detected in the external solution; and the cells themselves are easily removed by centrifugation.

There are three general ways in which hæmolysis may be brought about. In the first, the cells are placed in a solution which has a salt concentration less than that of the plasma; water passes into them and they swell up and finally burst. In the second, the cell membranes are damaged by means of a solution of ether, saponin, bile salts or other surface active substances. In the third, the blood is repeatedly frozen solid and thawed as rapidly as possible.

When placed in a hypotonic salt solution, the red cells, owing to their biconcave shape (Chapter 3, p. 69), can increase in volume very considerably before the surface membrane becomes stretched. The

amount of further swelling which can take place without rupture of the membrane and hæmolysis, varies from one cell to another, even in the same sample of blood from the same individual person ; in normal human blood, while the great majority of cells are just hæmolysed in about 0·4 per cent. NaCl, a small proportion (5 to 10 per cent.) burst in stronger solutions, and about the same proportion are unaffected unless the concentration is reduced to less than to 0·4 per cent. In this way a *resistance* or *fragility curve* can be constructed, showing the proportion of corpuscles hæmolysed, against the concentration of the solution in which they are placed.

The proportion of cells with small resistance is greatly increased in the blood of patients suffering from certain diseases such as familial hæmolytic jaundice ; when these patients have their spleens removed, their red cells attain a resistance which is normal, or even greater than normal. During recovery from anæmia (*e.g.*, pernicious, or after hæmorrhage), on the other hand, the proportion with a large resistance is greatly increased. The resistance to hæmolysis of a given cell probably depends upon its age, the membrane becoming steadily weaker during its life in the circulation until it finally gives way altogether.

The surface active substances do not necessarily destroy the cell membrane entirely, but make it permeable to sodium and potassium—perhaps only in certain places. The concentrations of sodium, potassium, chloride and bicarbonate then become more nearly the same inside and outside the cells, and there is an excess osmotic concentration of hæmoglobin within the cells : these then take up water, swell, and eventually burst, allowing the hæmoglobin to escape.

Interchanges across the Walls of the Capillaries—the Formation of Tissue Fluid and Lymph. The walls of the capillary blood vessels are very much more permeable to dissolved substances than are the cell membranes. Of the substances ordinarily present in the extracellular fluid all, except the plasma proteins, can penetrate freely and will diffuse to and fro between the blood and the interstitial fluid. Now the plasma proteins, like any other substances in solution, will contribute to the total osmolar concentration of the plasma. This, therefore, will be a little greater than that of the interstitial fluid, even though the concentrations of the crystalloidal substances are very nearly the same in both. Water will tend to pass into the plasma, accompanied by those substances to which the walls of the capillaries are freely permeable. As already mentioned, however, the tendency of water to move from solutions whose osmolar concentration is small to those whose concentration is large, may be opposed by the application of a hydrostatic pressure. Such a pressure, however, cannot be effective in practice unless there is a rigid membrane separating the two solutions concerned and impermeable to some, at least, of the dissolved substances. If this pressure is just sufficient to prevent any net movement of the water, it is equal to that part of the total osmotic pressure which is contributed by those substances which cannot penetrate the membrane. In the blood vessels there is such a hydrostatic pressure, produced by the action of the heart, and necessary for the circulation of the blood. If the capillary blood pressure exceeds the osmotic pressure due to the

proteins (about 25 mm. Hg in mammals), fluid will pass out from the plasma into the interstitial fluid ; at the arterial ends of the capillaries the pressure is high enough (about 35 mm. Hg) for this to occur. But at the venous ends of the capillaries, the hydrostatic pressure (about 20 mm. Hg) is less than the osmotic pressure due to the proteins, and fluid will be drawn in again. The net difference between (1) the capillary pressure (less the hydrostatic pressure in the interstitial spaces), and (2) the osmotic pressure difference between the plasma and the tissue fluid, is known as the *effective filtration pressure.* If this is positive, fluid leaves the blood stream ; if it is negative, fluid enters.

This filtration and absorption can be seen in the capillaries of the frog's mesentery. The mesentery is observed under a binocular microscope ; a capillary with a rapid circulation is chosen and blocked by pressing a fine blunt glass rod on it by means of a micro-manipulator. In some cases the mass of corpuscles remaining in the capillary moves towards the block, indicating that filtration is taking place, and in some cases they move away from the block, indicating absorption. A very fine glass pipette is now inserted into the capillary, connected with a water manometer, and the capillary pressure is measured, the block being still in place. The results of a large number of such experiments are shown in Fig. 8. 3, in which the rate of passage of fluid through the capillary wall is plotted against the capillary pressure. It will be seen that, on the average, fluid passes out when the pressure is greater than 11·5 cm. of water, and in when it is less than this ; independent measurements of the colloid osmotic pressure of frog's plasma indicate that it normally lies between 10 and 12 cm. of water.

The same general relation between the rate of filtration or absorption of interstitial fluid, and the excess or deficit of the mean capillary pressure above or below the protein osmotic pressure, can be demonstrated in preparations of the hind-limbs of cats or dogs ; but the experimental procedure is more elaborate and the evidence less direct.

In all the organs of the body, therefore, there is a flow of interstitial fluid out of the capillaries in the parts near the arterioles, and back into the capillaries in the parts towards the venules. The fluid would be expected to collect in the tissue spaces, and build up a hydrostatic pressure, until the inflow and outflow are equal. But in fact, the pressure in the tissue spaces is not ordinarily more than 1 or 2 mm. of mercury : the return flow into the plasma is in general rather less rapid than the outflow from the plasma and the excess fluid—which is known as **Lymph**—is carried off by the lymphatic system. This consists of very thin-walled vessels provided with valves, which all run towards the thorax, those from the hind limbs, abdomen, left side of chest and left arm all joining together to form the *thoracic duct,* which empties into the venous system at the junction of the left internal jugular vein with the left subclavian vein. Fluid is propelled along the lymphatics partly by means of the pumping action of muscular movement, as in the veins, and partly by reason of the negative pressure in the thorax.

The walls of the capillaries are not completely impermeable to the plasma proteins. But since the proteins pass through very much less readily than the water and salts, they get left behind during the filtration of the interstitial fluid ; the more rapid is the filtration, the more,

proportionately, are they left behind. The interstitial fluid, therefore, usually contains some proteins, the concentration varying in different parts of the body, according to the capillary pressure and the permeability of the capillary walls for proteins. Both may vary from part to part and from time to time.

The protein which is filtered off with the interstitial fluid tends to

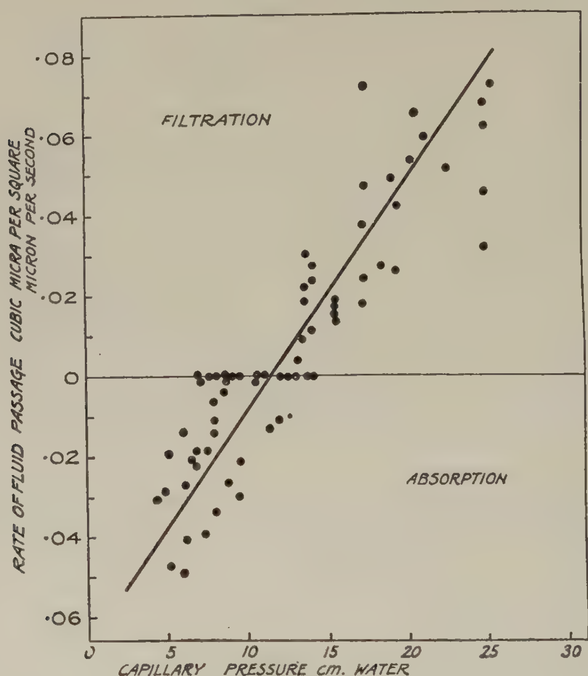


Fig. 8.3. The effect of the Pressure within a Capillary on the Passage of Fluid through its Walls.

Positive values of the rate of fluid passage indicate filtration of fluid ; negative values indicate absorption.

The observations were made on several different frogs, and the straight line, representing the average, passes through the line of zero flow at a hydrostatic pressure of 11·5 cm. of water (about equal to the colloid osmotic pressure of the plasma) ; fluid passes neither in nor out at this pressure, and the rate of flow is directly proportional to the pressure both above and below it. (Landis.)

be left behind again when the fluid is reabsorbed into the venous capillaries. Lymph, therefore, has a higher concentration of protein than has the fluid filtered off from the capillaries, the actual value being very variable (1 to 6 per cent.) ; lymph from the intestines and liver has a higher protein concentration (6 per cent.) than that from the muscles (2 per cent.). Otherwise lymph contains the same substances as does the plasma, and in about the same concentrations. It usually clots if left to stand, but not so rapidly as the blood, owing to the absence of

platelets. After a meal there is often a considerable quantity of fat present, giving it a milky appearance ; we have seen in Chapter 5 that the lymphatic system is the chief route for the absorption of fats from the alimentary canal. It normally contains, also, considerable numbers of lymphocytes.

Factors affecting the Volume of the Interstitial Fluid and the Flow of Lymph. Œdema. In the steady state, as already described, any excess of interstitial fluid is carried away in the lymphatics : but any change in the rate of lymph flow is ordinarily associated with an inverse change in the volume of fluid in the interstitial spaces. If, in any organs or tissues, there is a large increase in the rate of production of interstitial fluid, or a large decrease in the rate of its removal, these organs or tissues become swollen and puffy, and a condition of *œdema* is said to have developed. Such an accumulation of interstitial fluid in any part of the body is likely to occur, of course, if there is any obstruction to the lymphatic drainage from that part. But it will occur also, more generally, in response to any changes in the difference between the rate of filtration out of the capillaries and the rate of absorption into them. The factors which influence this may be grouped under three headings : (1) the capillary pressure ; (2) the difference between the osmolar concentration of the plasma and that of the tissue fluid ; and (3) the permeability of the capillary wall.

(1) Any rise in *capillary pressure* occasioned either by dilatation of the arterioles or obstruction of the veins, increases the rate of production of tissue fluid. Such a rise is a common accompaniment of activity in any organ, as has been described in Chapter 2 (p. 39).

(2) Two factors come into the consideration of the *osmolar concentration* : (*a*) the concentration of the colloids, and (*b*) the concentration of crystalloids.

(*a*) The concentration of the plasma proteins, and hence the colloid osmotic pressure of the plasma, can be reduced to a very low value by repeatedly bleeding an animal and re-injecting the blood corpuscles suspended in physiological salt solution ; the plasma proteins are thus removed, but not the corpuscles, so that the blood can still carry oxygen and carbon dioxide in a more or less normal manner. This process leads to a large increase both in the volume of the interstitial fluid and in the rate of flow of lymph. Such a condition is, of course, highly artificial.

(*b*) If the concentration of crystalloids is suddenly increased by injecting a hypertonic solution intravenously, the absorption of tissue fluid is increased. Although the crystalloids can pass through the walls of the capillaries quite rapidly, water can pass through even more rapidly. Equalisation of the crystalloid concentration in the blood and the interstitial fluid is brought about by a simultaneous passage of water into the blood and passage of crystalloids out of it. The flow of lymph practically ceases for a short time while this is occurring. The total volume of the blood is increased as a result of the inflow of water and thus the concentration of the proteins, and their osmotic pressure, is

decreased ; the rate of filtration is increased, the lymph flow starts again and continues at a rate larger than normal until the excess fluid has been removed. The net result, therefore, of introducing a hypertonic salt solution is the same as that of removal of the proteins : both act as *lymphagogues* (from " lymph," and the Greek word for " to lead "). The excess fluid is not all returned to the blood by the lymph, however, but mainly stays in the tissues until it is excreted by the kidneys.

The action of the crystalloid concentration is of importance, also, during the activity of any organ. The metabolites produced by the active cells will diffuse out into the tissue spaces, raise the osmotic pressure of the interstitial fluid and draw fluid from the capillaries. This, together with the rise in capillary pressure, accounts for the mild degree of œdema that often occurs in muscles after exercise.

(3) When a capillary dilates excessively, it allows more fluid to pass through its walls at a given effective filtration pressure and holds back protein less completely. Whether the physiological dilatation of the capillaries that occurs during activity of the organ supplied by them plays a part in the increased flow of lymph, is disputed, but it is of importance in connection with certain abnormalities. Histamine is one of the most powerful capillary dilators known, and it also injures the capillary endothelium, making it more permeable to large molecules. Histamine, therefore, increases the rate of formation of tissue fluid and of lymph. Its introduction below the surface of the skin through a needle prick results in the formation of a wheal, just as if the skin had been burnt. Similar reactions result from contact with the skin of the leaves of the poison ivy, or the stings of stinging nettles or jelly-fish ; these are known to inject small quantities of histamine (among other substances) into the skin. The release of histamine, or of some substance with similar actions, by damaged tissues will be referred to again in a later chapter, in relation to the " immunity reactions."

The increased permeability of damaged capillaries has been observed by the method described above (Fig. 8. 3). When alcohol or mercuric chloride was added to the blood flowing through them, fluid passed out much more rapidly, and there was practically no reabsorption except at very low pressures; they had thus become almost completely permeable to proteins.

There are several pathological conditions in which the rate of elimination of fluid is diminished, and the general water balance of the body disturbed (Chapter 7, p. 222) : an excess of fluid accumulates, and this is held chiefly in the interstitial spaces, giving rise to œdema. This occurs not only when the excretion of water fails, but also, and more commonly, when the excretion of salts fails : the control systems which preserve the constancy of the osmolar concentration of the body fluids bring about a simultaneous retention of water. An inadequate elimination of fluid will occur, most obviously, when there is a failure of the kidneys themselves, due to disease (nephritis) : but œdema, also, not infrequently accompanies disease of the heart. Heart failure is likely to produce a rise in venous pressure, as discussed in Chapter 1 ; but the consequent direct action on the rate of filtration from the capil-

laries is only partly responsible for the salt retention and œdema, the origin of which has not yet been explained.

Œdema also occurs in cases of severe malnutrition, when the supply of protein in the diet is grossly inadequate (*hunger œdema*). The plasma proteins are among the last to be sacrificed for the supply of energy or nitrogen to the rest of the body, so that the œdema is not due, to any considerable extent, to a simple reduction of the colloid osmotic pressure of the plasma. But the proteins of the less essential tissues are broken down to supply energy and to replace proteins unavoidably lost from the more essential tissues ; the fluid in which they were previously held thus becomes, as it were, " surplus to requirement." For some reasons, which are not obvious, this fluid is not completely eliminated, and accumulates in the interstitial spaces.

Physiological Saline Solutions

A great deal may be learnt about the way in which the various organs and tissues of animals perform their functions, by removing them from the whole animal and examining their behaviour in isolation. They will not then be surrounded by their normal interstitial fluid and this must be replaced by some artificial solution ; the substances which must be present in such a solution have been discovered largely by trial and error.

Owing to its automatic rhythmic activity, the heart is the most convenient indicator of responses to changes in its perfusion fluid, and the classical experiments of Ringer on this organ form the basis of our knowledge of the subject. If a frog's heart is isolated and placed in a solution, of whatever composition, of which the total osmolar concentration is substantially different from that of the blood, it soon ceases to beat. The requirement that solutions outside cells should be isotonic with the intracellular fluid has already been discussed, and is made obvious by the hæmolysis of the red blood cells in hypotonic solutions. That the solution should be isotonic is not, however, sufficient. The heart will not beat in solutions of non-electrolytes, such as glucose. It will do so, however, for a little time in an isotonic solution of sodium chloride (0·65 per cent. for the frog) ; but soon the beats cease, and the heart remains relaxed. If, now, a small amount of calcium chloride is added to the solution, the beats begin again ; but after a short while, the relaxation after each beat becomes progressively less complete, until, at last, the heart remains fully contracted, and ceases to beat. If at this stage a suitable small amount of potassium chloride is added to the solution, contractions begin again, and the heart may continue to beat fairly normally for many hours. These observations are illustrated in Fig. 8. 4. It is clear, therefore, that at least three salts must be present in an adequate physiological solution—sodium, potassium, and calcium chlorides—and the heart survives longer still if the solution is made slightly alkaline by adding sodium bicarbonate. Ringer, in his series of experiments, found the concentration of each of these salts which favoured longest survival of the heart-beat of the

frog and tortoise. Locke, working with isolated mammalian hearts, found the addition of glucose an advantage, and the compositions of these " physiological fluids " are given in Table 8. 2. The addition of a small quantity of a magnesium salt (0·01 per cent. $MgCl_2$, for example) is beneficial in experiments with many kinds of mammalian tissue (as in Tyrode's, and in Krebs', modifications of Ringer's solution), but has little effect on the heart. The concentrations of potassium and calcium may be varied slightly to suit the particular tissue in use.

TABLE 8. 2

Compositions of Blood Plasma and of Physiological Saline Solutions

	Ringer's Solution. (Frog's heart.) gram.	*Frog's blood plasma.* gram.	*Locke's Solution.* (Mammalian heart.) gram.	*Mammalian blood plasma.* gram.
NaCl	0·65	0·55	0·9	0·7
KCl	0·014	0·023	0·042	0·038
$CaCl_2$*	0·012	0·025	0·024	0·028
$NaHCO_3$	0·02	0·1	0·02	0·23
NaH_2PO_4*	0·001	0·02	—	0·036
Glucose	—	0·04	0·1–0·25	0·07
Water	to 100	(100)	to 100	(100)

* The weights given refer to the anhydrous salts. Appropriately greater weights of the hydrated forms, which are in common use, should be employed in making up Ringer's solutions.

On the whole, therefore, the necessary composition of an artificial environment for contractile (and other) tissues, as discovered empirically, is very similar to that of the interstitial fluid ; this, of course, is hardly surprising. Ringer's, and Ringer-Locke's solutions contain smaller concentrations of sodium bicarbonate, and in compensation, greater concentrations of sodium chloride, than do the corresponding frog's plasma and mammalian plasma, respectively. Such differences are necessary because the artificial solutions are in equilibrium with room air containing only very small amounts of carbon dioxide, whereas the plasmas contain much larger quantities of free carbon dioxide, mammalian arterial plasma, for example, being in equilibrium with 5 per cent. of carbon dioxide. The acidity of the solution is related to the ratio of the concentration of carbon dioxide to that of bicarbonate (Chapter 3, p. 84) ; reduction in the former, therefore, must be compensated by a reduction in the latter if the solution is not to be too alkaline. In some kinds of experiment, it is possible to keep the artificial saline solutions in equilibrium with 5 per cent. of carbon dioxide (usually in pure oxygen) without great inconvenience ; the bicarbonate concentration is then increased, and the chloride concentration decreased, to about the values found in the plasma (as in one form of Krebs-Ringer

solution). Alternatively, the bicarbonate and carbon dioxide may be omitted and replaced by an equivalent concentration of a sodium phosphate buffer mixture adjusted to pH 7·4 (as in the other form of Krebs-Ringer solution).

The subject of the action of individual ions on contractile tissues is inevitably a confused one, owing to the fact that different tissues respond in different fashion to excess or deficiency of any particular ion. There are, however, certain general rules. (1) Sodium salts occupy a unique position. The amount present in Ringer's solution is greatly in excess of the minimal amount necessary, muscles only ceasing to contract when more than 9/10 is replaced by some non-toxic substance such as glucose or sucrose. (2) The concentrations of calcium and potassium salts are interdependent, and must be in about the right ratio, though the absolute values may vary considerably. (3) Anions are relatively unimportant so long as they are not toxic ; chlorides are

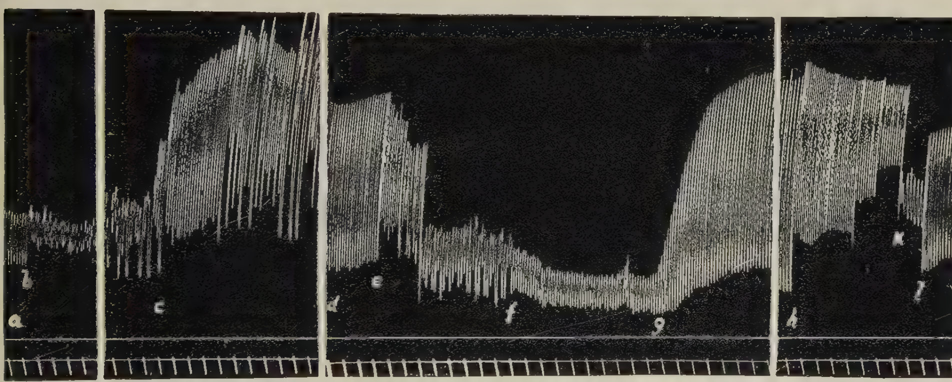


Fig. 8. 4. **The Action of Electrolytes on the Heart of the Tortoise.**

a–b : Beat of excised heart before perfusion.

b–c : Perfusion with *sodium chloride* solution, 0·75 g/100 ml. with sodium bicarbonate 0·01 g/100 ml.

c–d : Addition of *calcium chloride* (3 ml. of 0·1 M to 100 ml. of solution of sodium salts). Contraction improved but relaxation incomplete.

d–e : Addition of *potassium chloride* (6 ml. of 0·1 M to 100 ml. of solution of sodium and calcium salts). Relaxation improved and beats more regular.

e–f : Perfusion with solution of *sodium* salts again.

f–g : Addition of *potassium chloride* (6 ml. of 0·1 M to 100 ml. of solution of sodium salts). Beats more regular but contractions small.

g–h : Addition of *calcium chloride* (3 ml. of 0·1 M to 100 ml. of solution of sodium and potassium salts). Contraction greatly improved, with good relaxation.

h–k : Addition of excess calcium chloride. 3 ml. of 0·1 M solution added to 100 ml. of solution of sodium, calcium and potassium salts at each step in the record. Progressive impairment of relaxation.

k–l : Potassium chloride added in amount to correspond to the calcium chloride present—*i.e.* 2 ml. 0·1 M KCl for each 1 ml. 0·1 M $CaCl_2$. Improved relaxation but poor beat.

l onwards : Perfusion with normal Ringer's solution.

The smallness of the beat initially is due to the action of acid metabolites, which are formed during the anoxæmia accompanying dissection ; they are rapidly removed by the perfusion fluid. Recording : systole is upstroke.

(From Bayliss's "Principles of General Physiology," Longmans, Green & Co.)

usually employed, as these salts are all soluble, but the presence of sulphates, bromides, nitrates, bicarbonates, etc., has little influence on the survival of activity of an isolated tissue. (4) The hydrogen ion concentration must be not far removed from that of a neutral solution. Most tissues are favoured by a hydrogen ion concentration slightly on the alkaline side of neutrality. Slight acidity produces slowing or arrest of the heart-beat, relaxation of the tone of most unstriated muscles and phenomena analogous to fatigue in striated muscles.

If a muscle or nerve is placed in a solution which contains too high a concentration of potassium ions (say Ringer's solution modified to contain about three times the normal amount of potassium chloride), it becomes inexcitable ; the excitability can be restored by washing it with ordinary Ringer's solution. This effect accounts for the practical necessity of washing isolated muscles and nerves with Ringer's solution ; this removes the excess of potassium salts which diffuse out of those few fibres which are injured during dissection.

These, and many other kinds of experimental study make it clear that for proper functioning of the cells, the electrolyte composition of the fluid bathing them must be quite different from that of the fluid within them. The cells contain potassium ions and practically no sodium ions, while the fluid outside them must contain sodium ions and very little potassium ions. Many cells, also, contain practically no chloride, while the outside solution ordinarily contains chloride, although this can be replaced by other univalent ions.

The Cerebro-Spinal Fluid

There are four membranes covering the central nervous system. They are (1) the pia mater, which closely invests the nervous substance and carries the blood vessels to it ; (2) the arachnoid, separated from the pia by the sub-arachnoid space, which contains the *cerebro-spinal fluid ;* (3) the meningeal layer of the dura mater ; and (4) the periosteal layer of the dura mater. The venous sinuses are situated between the two layers of the dura and delicate processes known as the *arachnoid villi* arise from the arachnoid and penetrate into the venous sinuses. The relations between these structures are essentially similar in the skull and in the spinal column ; but since the pia ends, with the spinal cord, at the first lumbar vertebra, and the dura and arachnoid extend as far as the second sacral vertebra, the sub-arachnoid space is of considerable size in this region. Cerebro-spinal fluid can, therefore, be readily obtained by inserting a hollow needle (usually between the fourth and fifth lumbar vertebræ) into this space ; this is called *lumbar puncture*, and the fluid usually emerges at a rate of about one drop per second.

The cerebro-spinal fluid is formed by the plexuses of blood capillaries, known as the *choroid plexuses*, in the ventricles of the brain. This can be shown by the facts that (*a*) fluid can be collected from a tube inserted into either of the ventricles ; (*b*) blocking the outflow of a ventricle leads to its distension (hydrocephalus) ; and (*c*) if the choroid plexus is first removed, subsequent blocking of the outflow no longer

leads to a distension of the ventricle. From the ventricles, the cerebro-spinal fluid passes to the *cisterna magna*, an enlargement of the sub-arachnoid space at the base of the brain, and thence out and up in the sub-arachnoid spaces around the cerebellum and the cerebral cortex. It returns to the blood by absorption into the venous sinuses through the arachnoid villi. Flow up and down the spaces in and around the spinal cord probably results chiefly from changes in pressure in different parts of the system produced by movements of the head, or sudden changes in blood pressure.

The composition of the cerebro-spinal fluid is, in most respects, not far from that of an ultra-filtrate of plasma, containing a little protein (0·03 per cent.), and is thus very similar to that of the extracellular fluid shown in Fig. 8. 1. The actual concentrations of the various constituents are subject to considerable variation, as are those of the blood plasma ; but the changes in the two fluids tend to run hand in hand, so that the values of the concentration *ratios* remain relatively constant. Careful and precise analyses of the fluids drawn from one and the same animal show, however, that the values of these ratios are not quite those to be expected if the fluid were an ultra-filtrate ; the choroid plexuses must be regarded as secretory organs, and not merely as membranes which act as simple filters impermeable to proteins. Chloride ions, and to a less extent sodium ions, appear to be actively secreted into the cerebrospinal fluid, while potassium ions, urea and glucose are kept out. (The relatively low value of the glucose concentration may be due merely to the fact that it is metabolised by the tissues of the brain, and cannot diffuse in from the plasma very rapidly.)

That the structures which divide the cerebro-spinal fluid from the blood plasma have properties unlike those of the walls of the capillary blood vessels, is shown also by the rate at which substances pass across them. Substances with relatively small molecules pass very rapidly through the walls of the capillaries in muscles, for example, and their concentrations in the plasma and the interstitial fluid become sensibly identical in a matter of seconds. Passage from the plasma to the cerebro-spinal fluid occurs very much more slowly : it takes several minutes for the most rapid of the substances studied, ethyl alcohol, to attain equality of concentration ; and after 4 hours the concentration of creatinine in the cerebro-spinal fluid is only about 1/10th of that in the plasma. We are thus led to the idea of the *blood—cerebro-spinal fluid barrier* which controls the passage of substances from one to the other. The rate of penetration of this barrier, moreover, depends markedly on the lipoid solubility of the substance used, so that in its properties it resembles the cell membranes more than the capillary walls. There is also the *blood—brain barrier* between the plasma and the interstitial fluid of the brain, with properties similar to those of the blood—cerebro-spinal fluid barrier. Between the interstitial fluid and the cerebro-spinal fluid, however, the barrier is less restrictive and substances pass from one to the other relatively freely. It is difficult to get penicillin, for example, into the interstitial fluid of the brain in a useful concentration

by injecting it into the blood ; it is relatively easy to do so by injecting it into the cerebro-spinal fluid.

The total volume of the cerebro-spinal fluid in man is about 140 ml., and its rate of formation is about 0·5 ml. per minute. The pressure, measured by lumbar puncture, is about 150 mm. water when the subject is lying down, and rises to about 280 mm. water when he sits up. All these values may vary within wide limits.

Lumbar puncture is a procedure of considerable medical importance, for not only may it relieve increased intra-cranial pressure by providing an escape for excessive cerebro-spinal fluid, but it may yield valuable information enabling a distinction to be made between those diseases which exhibit characteristic changes in the nature and amount of the fluid. Inflammation of the membranes of the brain (meningitis), for example, is accompanied by an abnormally high rate of production of the fluid, by an increase in its protein content (which may be tenfold), by its composition approaching that of an ultra-filtrate of the plasma (except that the glucose concentration falls), and by an enormous increase in the content of leucocytes (Chapter 21). Normal fluid contains about 1 to 5 lymphocytes per cubic millimetre ; meningitic fluid contains hundreds of cells per cubic millimetre, which are mainly neutrophil leucocytes or lymphocytes, according as the infection is septic or tuberculous.

Local anæsthetics may be injected by lumbar puncture into the sub-arachnoid space to produce spinal anæsthesia. Air, or lipiodol (a heavy liquid opaque to X-rays) may be injected into the cisterna magna and the site of a tumour or inflammatory adhesion may then be discovered by the use of X-rays.

If the volume of the brain increases, as by the formation of a cerebral tumour, the intra-cranial pressure rises. This is transmitted to all points within the rigid brain-case by the cerebro-spinal fluid, with the result that the veins are compressed and the capillary pressure rises. The first structure to suffer is the optic nerve, and the obstruction of the veins running along it causes œdema of the optic disc, which can be observed with an ophthalmoscope ; this condition is known as *papillœdema*. The intra-cranial pressure is also increased in the condition known as *hydrocephalus*, which results from an excessive accumulation of cerebro-spinal fluid. This may be caused by obstruction to the fluid pathway from ventricles to sub-arachnoid space, or by blocking or maldevelopment of the arachnoid villi.

Rise in the pressure of the cerebro-spinal fluid is also caused by injections of isotonic or hypotonic saline into the blood, showing that its production does depend to some extent upon physico-chemical factors. Injections of hypertonic saline, on the other hand, cause a fall in the pressure, probably due chiefly to an osmotic withdrawal of fluid from the brain itself, with consequent diminution in volume. This is a procedure commonly used in brain surgery, where it is often advisable to reduce the volume of the brain before the skull is opened.

CHAPTER 9

URINE

HUNGER and appetite largely control the intake of food; thirst influences the intake of water; taste, to some extent controls the intake of salt. So might the water content and composition of the body be kept fairly steady, were it not that we like to eat different kinds of food, giving rise to different kinds and quantities of metabolic end-product and affecting differently the water balance of the body, and were it not for extraneous influences such as the social habits of drinking partners and the culinary customs of the home. The ample variation possible in urine formation provides for the spill-over of excess water and salts and indeed of most other soluble constituents of the plasma; so the constituents of the body fluids are kept at the steady levels required for normal physiological working as described in Chapter 8 on Body Fluids.

The daily output of urine varies widely in amount and composition; 1,500 ml. may be taken as representative for a man under average conditions in this country. Over shorter periods of time, the *rate of production* of urine varies between about 0·3 ml./min. and about 20 ml./min. according to the state of hydration of the body (1 ml./min. is equivalent to 1,450 ml. per day). Normally, urine contains about 4 per cent. solids, many of which are included in Table 9.1; but the amount varies inversely with the rate of water output, and the *total* (*osmolar*) *concentration*, as measured for example by the freezing point, varies from about one-sixth to four times or more that of plasma. The specific gravity, a rough measure of the total concentration, is usually between 1·015 and 1·025, but may fall to 1·002 or rise to 1·030. The *colour* of the urine, due to the presence of urochrome, a pigment of uncertain origin chemically related to hæmoglobin, is also a rough indication of the concentration. Urine is normally somewhat acid, its *p*H being about 6; this may vary between the limits of 4·7 and 8·2 according to the nature of the food and the amount of acid or alkali being excreted.

Quantitatively, the chief constituents of the urine are urea and the chlorides, sulphates and phosphates of sodium and potassium (Table 9. 1). A concentrated urine often deposits amorphous sodium and potassium urates on cooling. The precipitate is coloured pink by uro-erythrin and can be redissolved by warming. Another amorphous deposit may occur in normal urine, namely, phosphates of the alkaline earths; these have a low solubility in alkaline solution, and are precipitated (*a*) (usually as $Ca_3(PO_4)_2$) when the urine is alkaline when voided, or (*b*) (as NH_4MgPO_4—" triple phosphate ") when previously acid or neutral urine becomes alkaline on standing owing to bacterial conversion of

the urea into ammonium carbonate. The phosphate is dissolved by the addition of dilute acetic acid. Crystalline deposits are usually associated with abnormal processes. In acid urine, calcium oxalate, cystine, leucine or tyrosine may be found, whereas in alkaline urine, calcium carbonate and phosphates are the commonest.

Urea is the chief nitrogenous end-product of protein metabolism and occurs in large quantities in normal urine. The ability of a kidney to concentrate urea is considered a valuable index of functional activity in disease. It is estimated by administering 15 g. of urea in 100 ml.

TABLE 9. 1

Typical Concentrations in Man

	Plasma g. per 100 ml.	Urine g. per 100 ml.	Urine/Plasma Concentration ratio
Water	90–93	96	1·05
Proteins	7–9	0	0
Urea	0·03	2	60
Creatinine	0·001	0·15	150
Uric Acid	0·002	0·05	25
Glucose	0·10	0	0
Sodium	0·32	0·35	1
Potassium	0·02	0·15	7
Calcium	0·01	0·015	1·5
Magnesium	0·002	0·01	5
Chloride	0·37	0·6	2
Sulphate	0·003	0·18	60
Phosphate	0·003	0·12	40
Ammonia*	0·0001	0·04	400

* Ammonia is synthesised in the kidney.

of water by mouth and collecting urine at hourly intervals ; the first sample may be dilute owing to diuresis, but less than 2 per cent. in the second sample would be unusual in normal kidneys.

Other nitrogenous substances in urine occur only in relatively small amounts. Ammonium salts, usually in small concentration, may be increased in acidosis. A large proportion of the ammonia in normal urine is formed in the kidney itself. Uric acid is formed from nucleins, and about one-half persists in starvation and is therefore regarded as of endogenous origin ; the other half varies in amount with the diet, and is of exogenous origin. Creatinine in the urine is probably formed mainly from the creatine in muscle.

When urine has been allowed to stand for some time, a faint cloud of mucus from the walls of the bladder and urinary passages can often be seen.

Abnormal constituents appear in urine from normal kidneys when soluble and diffusible foreign substances, including many drugs, have been administered. Or again, they appear when certain normal constituents of plasma are present in excessive concentration. For example, the concentration of glucose, normally around 0·10 per cent. in plasma, may rise as a result of swallowing 200 g. or more within a short time ; if the plasma concentration then exceeds a " threshold value " of about 0·18 per cent., some of the glucose is excreted in the urine, a finding known as " alimentary glycosuria." Likewise in some diseases, normal kidneys will eliminate substances not normally secreted, for example, glucose and ketone bodies in diabetes mellitus or bile salts and bile pigments in jaundice. Diseased kidneys may allow substances which are normally retained in the plasma to pass into the urine, and most characteristically so, plasma protein. Proteinuria is characteristic of nephritis, of failure of adequate blood supply to the kidney and of the action of certain poisons on it. A transient appearance of protein occasionally occurs in the urine of healthy people, especially adolescents, after severe exercise and after prolonged standing ; the latter may be due to a rise in pressure in the renal vein. Albuminous casts of the tubules may appear in the urine secreted by diseased kidneys, and cells derived from the blood or from the excretory organs are present in the urine in certain pathological circumstances.

The Structure of the Kidney

Urine formation from plasma cannot be understood without reference to the structure of the kidney. The urine is formed in long unbranched tubules called *nephrons* many of which combine to form a smaller number of collecting ducts from which the urine is discharged into the pelvis and so to the single channelled ureter. Nephrons are very numerous, well over a million have been found in pairs of dogs' kidneys and there are probably two million or so in human kidneys. Each nephron begins with a wider blind end, the glomerular capsule (Fig. 9. 1) into which protrudes a bunch of blood capillaries known as the glomerular tuft or the glomerulus. The capillary walls are covered by a thin membrane forming the blind end of the tubule and called the glomerular membrane. This is continuous with the outer cup-shaped membrane (Bowman's capsule) which forms the wide end of the tubule, the space between the inner glomerular membrane and the outer funnel-like membrane being called the capsular space which contains glomerular fluid. The idea that the function of the glomerulus is ultrafiltration, that is transport of only the non-colloidal elements of blood into the tubule under influence of blood pressure, was first suggested by the microscopic appearance of Bowman's capsule but has since been proved experimentally.

Beyond the neck of the funnel, the thin flat cells of the outer capsule change into the columnar or cubical granular cells of the proximal convoluted tubule, resembling in appearance the cells in secretory organs. The glomerulus and proximal convoluted tubule are found in the outer zone, the cortex, of the kidney. The tubule then takes a dip into the medulla before returning to form the distal convoluted tubule near its own glomerulus ; the loop formed in this way is known as Henle's loop and the part it plays in concentrating urine by a counter-current mechanism has been discovered more recently. Beyond the

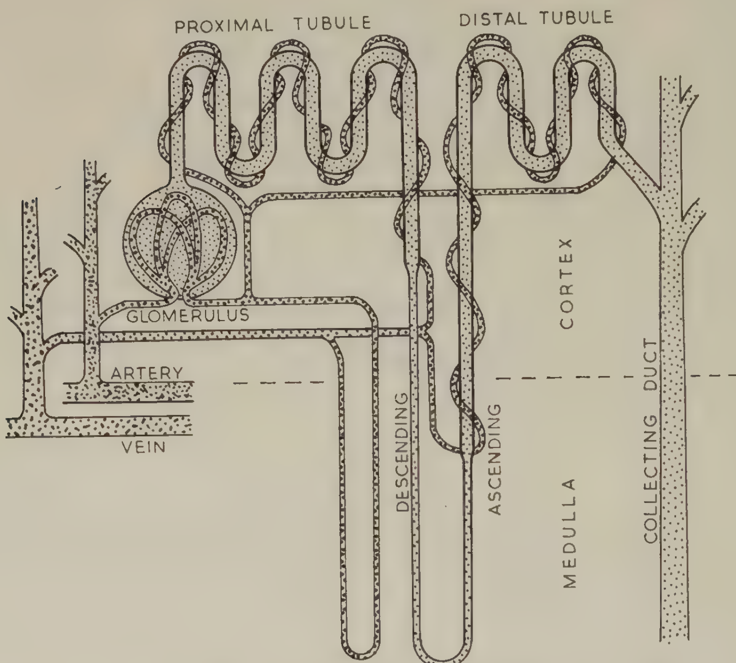


Fig. 9. 1. Diagram of a nephron and associated blood vessels.

The proximal and distal tubules have been reduced in length ; and the arrangement of these, as well as that of the dual system of capillaries— those within the glomerulus in series with those ramifying over the tubules— has been considerably simplified.

distal tubule, the tubule plunges once more toward the medulla to enter a collecting duct.

The *blood supply* to these renal elements is arranged in a peculiar manner. The glomerulus receives blood from branches of the renal artery by a short, wide, " afferent " vessel. From the glomerular capillaries the blood is collected into a longer and narrower " efferent " vessel, which, in turn, divides into capillaries distributed over the tubules. From the peri-tubular capillaries the blood collects into venous sinuses, where it reaches the radicles of the renal vein (Fig. 9. 1). In the mammalian kidney, nearly all the blood which reaches the peri-

tubular capillaries has previously passed through the glomerular capillaries. There are, however, shunts in this circulation which play a disputed part in normal function. Thus the blood normally approaches the glomeruli through the wide channels of the renal arterioles and vasa afferentia, and emerges through the narrow vasa efferentia to reach the tubular capillaries. Owing to these two anatomical factors—the double capillary system and the difference in diameter of the vasa afferentia and efferentia—the blood pressure in the glomerular capillaries is much higher than in capillaries in other organs. Its pressure in the peri-tubular capillaries, on the other hand, may be even lower than the normal capillary pressure elsewhere in the body. Thus, on histological grounds alone, it appears that blood at high pressure is separated from the lumen of the renal tubule only by the thin walls of the glomerular capillaries and by the single layer of flattened cells composing the inner membrane of Bowman's capsule. This is most aptly arranged to provide hydrostatic pressure for the filtration in the glomerulus mentioned earlier. The peri-tubular capillaries, on the other hand, may be at a lower pressure than capillaries in other organs, and this would facilitate the reabsorption of water and other substances which will be shown later to occur there.

The Secretion of Urine

The essential problem of renal physiology is presented in Table 9. 1, in which the constituents of urine and plasma are compared, and it is shown that some, such as proteins and sugar, are retained in the blood, and others, such as urea and sulphates, are largely concentrated in their passage from blood to urine. A third group of substances, such as sodium salts, appear in the urine in lower or higher concentration than in blood, according to the needs of the organism. In order to concentrate substances in solution an expenditure of energy is necessary. This is comparable with the energy necessary to concentrate a gas, as in pumping up a tyre. A source of physical energy, namely, the arterial pressure, is available in the kidney, and one can imagine a machine which would use this energy for the formation of urine. But our knowledge of the structure and mode of action of the kidney is not consistent with its acting only as such a machine ; for the most part, the urine is formed by a process of active secretory work, *i.e.* the performance of chemical work at the expense of energy derived from the metabolism of the cell. In the kidney the metabolism is mainly oxidative, and the kidney consumes oxygen at a high rate.

The oxygen consumption of the kidney is, however, not simply related to the osmotic work done in urine formation, for it is unchanged in diuresis due to administration of water or of " osmotic " diuretics, *e.g.* urea. The arterio-venous oxygen difference is lower, about 2 ml./100 ml., than in other organs, and instead of varying inversely with the blood-flow, as would be expected if the oxygen consumption were to remain relatively constant, it is practically unaffected by changes in blood-flow unless this is reduced to quite abnormally low values.

These unique features constitute important unsolved mysteries in renal physiology.

Theories of Renal Secretion. In 1842 Bowman, working in London, gave the first description of the main histological features of the kidney. From these he inferred that blood at a high pressure passing through the thin-walled glomerular capillaries would be likely to filter a watery fluid into the nephron. The tubules with cells resembling in appearance those in other glands with a well recognised secretory function, would be likely to add the main solid constituents of the urine to the watery fluids passing down from the glomerulus. In 1844, Ludwig in Germany proposed the *filtration-reabsorption theory* which differed from Bowman's view in supposing that a much larger volume of glomerular filtrate contained all the solid constituents subsequently appearing in the urine, the solids being concentrated during passage along the tubules by transfer of most of the water and some other substances to which the tubules were permeable back into the blood. Cushny emphasised the osmotic work performed by the tubule cells in concentrating the urine by reabsorption of water, and later Richards and his school in America used micropipettes to obtain samples of glomerular and tubular fluid for chemical analysis and thus produced the experimental proof that the filtration-reabsorption theory provides an adequate explanation of the formation of urine in so far as most of its normal constituents are concerned. Many foreign substances, such as phenol red or diodone, when injected into the blood stream appear in the urine, however, in much greater amount than can be accounted for by the filtration-reabsorption theory which confines tubular function to the withdrawal of substances from the lumen of the tubule and their return to the blood. It is now generally agreed that such foreign substances, if present, and a few normal constituents of urine, are transferred to the tubular urine direct from the blood in the peri-tubular capillaries by secretion into the lumen of the tubule. In the hands largely of Homer Smith and his school in America, the filtration-reabsorption theory, so modified, has led to the development of methods of assessment of the glomerular filtration rate and consequently of a quantitative determination of both the reabsorptive and secretory functions of the tubules.

Agreement on the main outline of the theory of renal secretion was reached in the late 'thirties after nearly a century of ingenious experimentation and often heated polemic. However, in the 'fifties, Wirz and his colleagues in Switzerland introduced the counter-current theory of tubular function which is widely accepted though not all its implications have yet been explored.

The word " secretion " is commonly used in rather a special sense in connection with the renal tubules ; both " reabsorption " and " secretion " are really forms of secretion in the general physiological sense ; reabsorption involves an indirect secretory activity of tubule cells whereby substances (*e.g.*, glucose) are removed from the fluid in the tubules, where their concentrations are low, to the blood where their concentrations are higher. Direct " secretion " involves transference of other substances (*e.g.*, diodone or para-amino-hippuric acid) from the blood, where their concentrations are low, to

the tubules, where their concentrations are higher. Both processes require the performance of chemical work.

Glomerular Function

In the past the glomeruli were thought of either as secretory organs, concentrating some of the constituents of urine, or as organs of filtration producing something less than the volume of urine to which the more concentrated constituents were added by secretion during its passage down the tubules. A third view, which has now established itself was that the glomerular fluid was filtered at, perhaps, a hundred-fold the

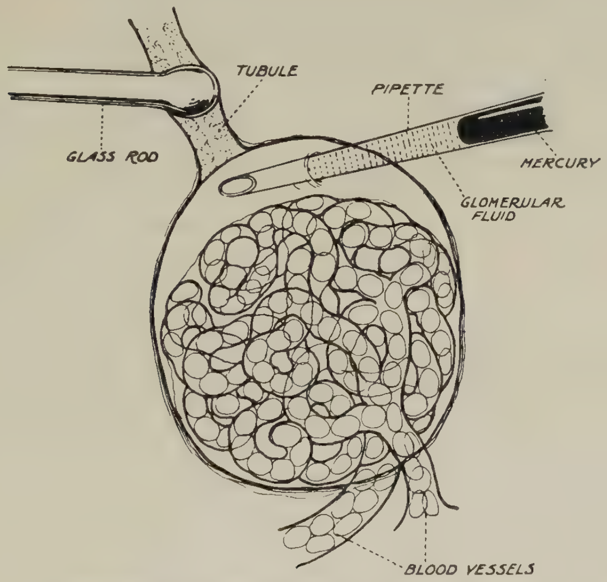


FIG. 9. 2. Diagrammatic representation of Richards' method of obtaining a sample of glomerular fluid from Bowman's capsule in the frog.

A very fine pipette (7 to 15μ diam.) connected with a reservoir of mercury and filled with mercury up to its tip, is thrust through the wall of the capsule, and fluid is withdrawn from the capsular space by lowering the mercury reservoir. The tubule is blocked by pressing on it with a fine glass rod, so that no fluid can be sucked back into the capsule.

rate of urine formation and that the fluid contained all or nearly all the solid constituents in urine, some of which were concentrated by the reabsorption of water during passage down the tubules. Complete experimental proof that glomerular fluid is formed from plasma by a physical process of filtration would be provided if it could be shown (1) that the composition of the fluid is exactly that of protein-free plasma, (2) that the glomerular membrane has pores of such size that plasma protein and larger molecules cannot pass but smaller molecules can pass, and (3) that the hydrostatic pressure-fall across the membrane is enough to drive the glomerular fluid at the known rate across the

membrane of known resistance. The first two lines of evidence have been provided by convincing experiments. The third is supported only by rather speculative calculations from indirect observations and it could still be argued that some active pumping assists hydrostatic pressure in driving fluid across the glomerular membrane. In the absence of evidence, this possibility is usually ignored.

Direct evidence that glomerular fluid has the composition of protein-free plasma was first produced in amphibian kidneys by A. N. Richards and colleagues by introducing the point of a micropipette (7 to 15μ diam.) into the capsular space and slowly withdrawing fluid from it (Fig. 9. 2). The glomerular fluid so obtained was almost free from proteins, and could not, therefore, be suspected of contamination with blood ; it contained chlorides and glucose even when these were absent from the bladder urine collected at the same time. Owing to the extremely small amounts available (1 cu. mm. at the most), special methods of chemical analysis had to be devised ; these have demonstrated the equality of the concentrations in plasma and glomerular fluid not only of all the chief normal constituents of urine such as urea, chlorides, and creatinine, but also of certain diffusible foreign substances, such as inulin. Here, then, is unequivocal evidence that in the amphibian kidney the glomerular membrane acts as a semi-permeable membrane, allowing crystalloids to pass through it, and preventing the passage of proteins. Similar observations on the composition of glomerular fluid were later performed on small mammals.

In larger mammals, such as the dog, evidence of filtration can be obtained by abolishing the tubular activity which normally converts glomerular fluid into urine. The liquid which emerges from the ureter has then practically the same composition as an ultrafiltrate of plasma. Such observations have been performed on isolated kidneys excised from dogs and perfused with blood from pump-lung circuits. Tubular activity can be removed reversibly by cooling the blood or less reversibly by poisons such as cyanide.

It may be argued that this experiment proves the existence of filtration in the cold kidney, but not in the warm kidney, for even if the glomerular membrane were secretory in function, cold would presumably abolish that secretion. But cold does not, as far as is known, increase the permeability of membranes ; consequently the filtration which takes place when the kidney is cold, must also take place when it is warm.

The second property of the mammalian kidney proving the existence of a filtration mechanism in the kidney is concerned with the nature of the filter. A filter implies a membrane which will allow the passage of particles below a certain size, but retain larger particles. An ultra-filter, such as the glomerular membrane, should allow the passage of molecules below a certain size, but retain larger molecules. The following table (after Bayliss, Kerridge and Russell) shows that the kidney differentiates between molecules of different sizes in just such a simple physical way. Fig. 9. 3 shows that the position of the filter in the kidney

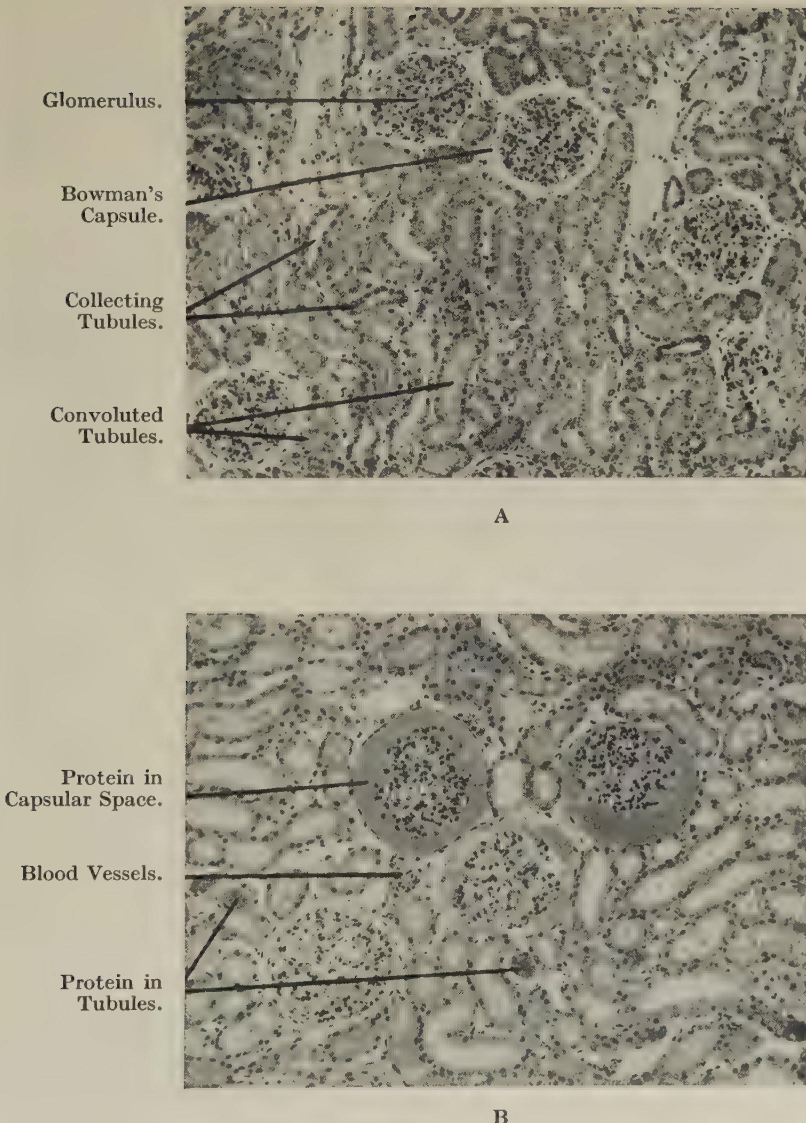


FIG. 9. 3. Sections through isolated and perfused kidneys of dogs. (Magnification × 140.)

A. Perfused with normal defibrinated blood for one and three-quarter hours.

B. Perfused with normal defibrinated blood for one and a half hours, and with blood containing egg albumin for half an hour. Note the presence of protein in Bowman's capsule, indicating that the glomerular membrane is permeable to proteins of relatively low molecular weight (less than 70,000). (Bayliss, Kerridge and Russell.)

TABLE 9. 2

		Molecular Weight
Proteins excreted		
Gelatin	about	35,000
Bence-Jones	,,	35,000
Egg albumin	,,	35,000
Hæmoglobin	,,	67,000
Proteins not excreted		
Hæmoglobin	about	67,000
Serum albumin	,,	72,000
Serum globulin	,,	170,000
Casein	,,	200,000
Edestin	,,	200,000
Hæmocyanin	,,	5,000,000

Hæmoglobin and serum albumin have molecules near the borderline and are believed normally to leak very slowly across the glomerular membrane, the amount that has leaked being reabsorbed by the tubules. If more than a small amount appears in the tubules, due to excessive concentration in the plasma or abnormal permeability of the glomerular membrane, the very limited reabsorptive capacity of the tubules will be exceeded and the protein will appear in the urine.

is in fact the glomerulus, since the protein with a sufficiently small molecule is shown to pass from the blood into the glomerular space.

The discrimination of the glomerulus between molecules of different size has also been shown in other series of compounds. For example, the dextrans, polysaccharides used therapeutically as blood substitutes, can be hydrolysed to any extent required to produce molecules of weights from that of glucose up to several millions, and the fractions of different molecular sizes have been separated and injected by Wallenius. In man and the dog molecules of dextran exceeding about 47,000 in weight fail to be excreted in the urine. Molecules with a weight of 5,000 to 6,000 or lower pass freely into the urine, whereas molecules of intermediate size pass more slowly. Human patients and animals with proteinuria allow the passage of larger dextran molecules (50,000 to 100,000) into the urine indicating an abnormally permeable glomerular membrane. Size as indicated by molecular weight is not the only property of molecules which affects their passage through membranes. The shape of the molecule and presence of electric charges are also concerned. Differences in these presumably account for the difference in the molecular weights of the proteins and polysaccharides which are just able to pass through the glomerular membrane.

The third line of evidence required to complete the proof of filtration across the glomerular membrane is a relation between hydrostatic pressure across it and rate of formation of glomerular fluid appropriate to a purely physical process. The glomerular filtration rate should be fairly linearly related to the hydrostatic pressure minus the colloid osmotic pressure, *i.e.* the effective filtration pressure. Curiously, this

evidence is far from manifest. In normal animals a large rise in arterial pressure, which might be expected to produce a corresponding rise in glomerular capillary pressure, raises the glomerular filtration rate very little. This is attributed to an altogether different mechanism discussed later under the heading "Autoregulation." In the isolated kidney of the dog, the control of autoregulation is less perfect than in the

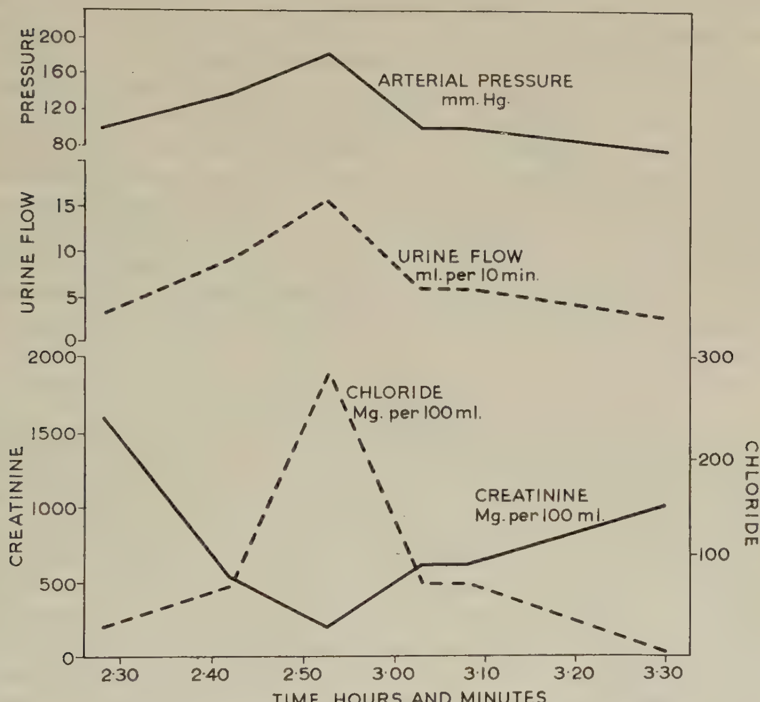


FIG. 9. 4. The influence of arterial pressure on the isolated kidney of the dog.

Increase of arterial pressure produces a large increase of urine flow, and a change in its composition (decrease of creatinine concentration and increase of chloride concentration) such that the concentrations of the solutes move in the direction of those in the serum. Serum creatinine 75 mg. per 100 ml. Serum chloride 760 mg. per 100 ml. (chloride estimated as NaCl). Temp. 37° C. (Gilson and Winton.)

normal kidney and the glomerular filtration rate follows the arterial pressure changes more appropriately. An experiment of this kind is depicted in Fig. 9. 4. How to calculate glomerular filtration rate from serum and urine creatinine concentrations and the urine flow will be described later under the heading "Clearances." An important property of the kidney is depicted in Fig. 9. 4, namely that an increase in urine flow is accompanied by a change in its composition in the direction of that of the plasma ; substances which are more dilute like chloride,

or more concentrated like creatinine, in the urine than in the plasma, become less so in diuresis.

The pressure fall across the glomerular membrane has been calculated to be about 55 mm. Hg in a dog kidney, more than half of which is devoted to overcoming the osmotic pressure of the plasma proteins. Under average conditions with mean arterial pressure of, say, 115 mm. Hg, the glomerular capillary pressure may be about 75 mm. Hg and the pressure in Bowman's capsule about 20 mm. Hg. Micropipette measurements indicate pressures of 2 to 3 mm. Hg in the distal tubules at resting urine flows but the pressure rises during diuresis.

Dilution of plasma proteins reduces the colloid osmotic pressure which itself opposes glomerular filtration. The glomerular filtration rate is, therefore, increased. This produces " dilution " or " saline diuresis," though this diuresis is augmented by an effect on the tubules. Glomerular capillary pressure may vary although arterial pressure remains unchanged. For example, caffeine increases glomerular pressure by dilatation of the preglomerular arterioles (vasa afferentia) whereas certain doses of adrenaline raise the glomerular pressure by constriction of the post-glomerular arterioles (vasa efferentia)—both drugs producing diuresis.

Plasma Clearances

The interest in plasma clearances is due to their value in assessing both renal blood flow and glomerular filtration rate in unoperated man and other animals. Moreover, they can discover substances which are excreted in greater or less amount than that present in the glomerular filtrate and must, therefore, be partly " secreted " or reabsorbed in the tubules.

By *plasma clearance* is meant the volume of plasma which contains the same weight of the substance as is contained in the volume of urine secreted in a minute. Take, for example, the values of creatinine concentrations in plasma and urine given in Table 9. 1. These would correspond to a resting urine flow of, say, 1 ml. per minute which, at the creatinine concentration of 0·15 per cent., would contain 1·5 mg. Since 100 ml. of plasma contained 1 mg. of creatinine, equal amounts (1·5 mg.) of creatinine would be found in either the 1 ml. urine or in 150 ml. plasma. The plasma clearance is then said to be 150 ml. per minute. The concentration in the urine divided by the concentration in plasma is called the *concentration ratio* or U/P ratio ; and this, multiplied by the urine flow (F), (*i.e.* $F \times U/P$) gives the plasma clearance.

Since equal volumes of glomerular filtrate and of plasma contain the same weight of a freely filtered substance, the *glomerular filtration rate* would equal the plasma clearance for a substance which is neither removed from, nor added to, during passage of the filtrate through the tubules. Creatinine appears to be such a substance in dogs, sheep, rabbits and some other animals. In them, the creatinine clearance is a measure of the glomerular filtration rate. In man, however, a little

creatinine is added to the filtrate by the tubule cells and inulin is generally accepted as the best substance for measurement of filtration rate.

The choice of inulin for this purpose is justified by the following considerations. Inulin solutions, if pure, can be injected intravenously into man and other animals without harm and without affecting the circulation or the activity of the kidney. Inulin is not metabolised in the body. It is freely filterable through the glomerular membrane since its concentration in frog's glomerular fluid is equal to that in the plasma. The belief that inulin is neither reabsorbed nor secreted in the tubules is mainly based on the following observations : (*a*) The molecular weight of inulin is 5,200, and being a long chain polysaccharide, it diffuses at only about double the rate of hæmoglobin. It is, therefore, unlikely to diffuse out of the tubules. (*b*) Inulin (mol. wt. 5,200) and creatinine (mol. wt. 112) have the same plasma clearances when measured at the same time in dogs and in many other species. It would seem that two substances with chemical natures and diffusion rates so widely different could not be treated exactly alike by tubule cells in either the process of secretion or of passive diffusion. (*c*) The inulin clearance is independent of its concentration in the plasma ; whereas the clearances of most other substances which are higher or lower than that of inulin when their concentrations in the plasma are low, approach the inulin clearance progressively as their concentrations in the plasma rise (see Fig. 9. 5). (*d*) In man and animals in which the creatinine clearance exceeds the inulin clearance, there are other substances, *e.g.* sorbitol and mannitol, whose clearances are identical with the inulin clearance. (*e*) Inulin cannot be excreted by certain fish whose kidneys contain no glomeruli (" aglomerular kidneys "). All constituents of the urine in these species must be secreted by the tubules and this emphasises the incapacity of tubules to secrete inulin.

On the basis of these considerations there is now widespread agreement that the glomerular filtration rate in the normal kidney may be measured in terms of the plasma clearance for inulin. In kidneys damaged by disease or poison, in which abnormally low inulin clearances are found, the possibility of abnormal permeability of the tubules leading to loss of inulin by passive reabsorption must, however, be borne in mind.

The inulin clearance in man ranges in healthy individuals from 90 to 170 ml. per minute—the average is about 125 ml. per minute for men and 118 ml. per minute for women—all reduced to the standard size of 1·73 sq. metres surface area, *i.e.* the area generally regarded as representative for men and women of average size. In a given individual, the inulin clearance is surprisingly difficult to change. In water diuresis, for example, an increase in urine flow from a resting value of 1 ml. per minute to nearly 20 ml. per minute which is about the maximum, involves little or no change in inulin clearance.

The renal plasma flow could be measured in terms of the plasma clearance if a substance could be found which is completely removed

from the plasma during a single passage from renal artery to renal vein. To obtain blood flow from plasma flow the hæmatocrit value would also be needed. Such a substance must be unable to diffuse out of the corpuscles during their passage through the kidney. With such an ideal substance, all of it would pass from the arterial plasma to the urine, leaving none in the venous plasma. Its *extraction-ratio* would be 1·0. The amount appearing in the urine in one minute would be the same as the amount in the arterial plasma passing through the renal artery in one minute. If its concentration in the plasma entering the kidney was also known, the volume of plasma traversing the kidney per minute could be calculated, that is, the plasma clearance would be equal to the plasma flow.

Substances approaching the required conditions are para-amino hippuric acid and the organic iodine compound known as diodone in England and iodopyracet in America. If the plasma concentration of diodone is below the threshold above which the clearance begins to fall, the renal venous plasma, measured by catheterisation of the renal vein (Chapter 2), is found to contain only about one-tenth of that in the arterial plasma. The kidney has extracted, say, nine-tenths of the diodone, and so long as this extraction ratio remains constant, the necessary correction can be made in calculating the renal plasma flow. The extraction ratio fortunately appears to remain fairly constant in normal kidneys. It falls so far, however, in kidneys damaged by disease or poisons that the method often becomes inapplicable.

In resting normal men, the diodone (iodopyracet) clearance at low plasma concentrations averages about 655 ml. per minute, corresponding to a renal plasma flow of about 720 ml. per minute and a renal blood flow of nearly 1,300 ml. per minute. Taking the resting cardiac output at 5,000 ml. per minute, it appears that at rest the supply to the kidneys accounts for about one-quarter of the output of blood from the heart. The proportion of plasma entering the kidney which is removed by filtration in the glomeruli is known as the *filtration fraction* ; this is 125/720, *i.e.* about one-fifth.

Tubular Function

Protein-free plasma, emerging from the glomerulus to enter the tubules at about 125 ml. per minute, is transformed into a very much smaller volume of urine leaving the further ends at, *e.g.*, 1 ml. per minute. The main secretory work of the tubules is done in concentrating substances, such as urea, which are present in large amounts and which are relatively highly concentrated in the urine. Important also is the reabsorption of substances valuable to the body, such as glucose, which are present in plasma but normally absent in urine.

The reabsorptive function of the tubules has been unequivocally demonstrated in the kidney of the frog and the guinea-pig by the experiments of A. N. Richards and his colleagues, already mentioned, in which a comparison of the composition of the glomerular fluid and bladder urine showed that, in suitable circumstances, glucose and chlorides

might be present in the former, but absent from the latter. Moreover, the quantity of the glomerular fluid which they collected in a given time, multiplied by the number of glomeruli, was much greater than the volume of urine which appeared in the same time. Water, therefore, is reabsorbed, consequently substances which are not reabsorbed must appear in the urine in a higher concentration than in the plasma.

In the isolated kidney of the dog the reabsorptive function of the tubules can be shown by abolishing secretory activity by cooling or poisoning. Cooling, for example, produces a two- or three-fold increase of water output and an enormous increase of chloride output. Moreover, glucose now appears in the urine in about the concentration in which it is present in the plasma. Cyanide poisoning has much the same effects. The secretory activity of the tubules, which has been abolished, must, therefore, have the effect of removing all the glucose, and much of the chlorides and water, from the glomerular filtrate and restoring them to the blood.

In man, 125 ml. per minute, that is, nearly 200 litres in twenty-four hours, are filtered in about two million glomeruli, producing about 1·5 litres of urine. On the average, therefore, each glomerulus filters 0·1 ml. a day, nearly all of which is reabsorbed in passage down its tubule, of average length about 5·5 cm., and with reabsorptive surface greatly increased by microvilli (brush border) in the proximal segment.

Most products of metabolism, which appear more concentrated in urine than in plasma, are so concentrated because they are less reabsorbed in the tubules than is the water. Nevertheless, some substances are certainly concentrated beyond this level by transfer from peri-tubular capillary blood to the lumen by " secretion " by the tubule cells. Most striking of these are the substances already mentioned whose plasma clearances approach the total plasma flow through the kidney. If nine-tenths of such a substance, say diodone, which reaches the kidney in the plasma is excreted in the urine, and if this were derived entirely from the glomerular filtrate, the plasma emerging from the glomerular capillaries would contain only one-tenth of its normal content of water and the blood would be much too viscid to pass through the vasa efferentia. In fact, only about one-fifth of the water in the plasma is removed by glomerular filtration, leaving four-fifths in the plasma in the vasa efferentia. Direct secretion of some substances into the lumen of the tubule is, therefore, certain. Among other such substances are penicillin and the dye, phenol red, which cannot readily pass the glomerular membrane because much of it is bound to plasma protein. Many drugs are similarly protein-bound. Creatinine in some species, such as man, is secreted in small amounts; in others, such as the dog, concentrated only by the reabsorption of water. The plasma clearances of all " secreted " substances are higher than that of inulin but approach this as their concentration in the plasma increases (Fig. 9. 5). This is because there is a maximum quantity per minute which the tubule cells can secrete known as the " transport maximum " or Tm ; whereas the amount contained in the glomerular filtrate per minute increases

directly with concentration in the plasma. The filtered portion, thus, increasingly dwarfs the secreted portion of the substance appearing in the urine.

Among substances which are reabsorbed, urea is quantitatively important. Its clearance is usually about one-half of that of inulin, and is independent of plasma concentration. The other half of the urea in the glomerular filtrate is passively reabsorbed from the tubule, the higher the urine flow the less, proportionately, being reabsorbed and the higher the clearance relative to that of inulin. Glucose is entirely

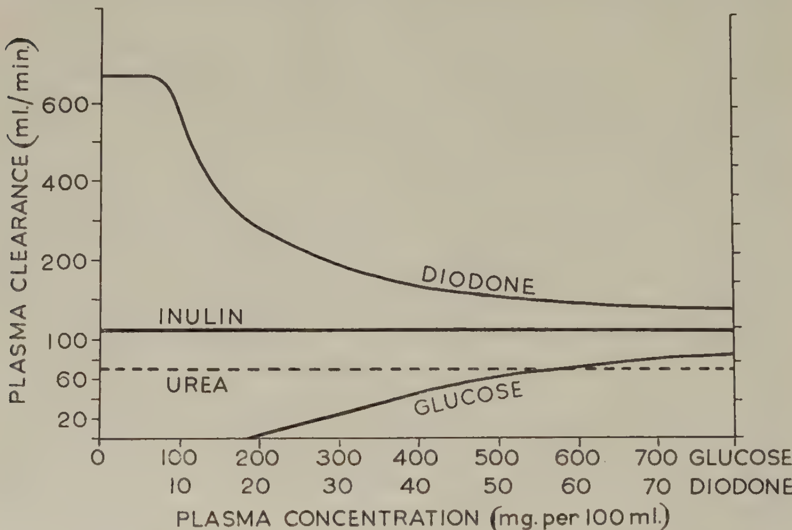


FIG. 9. 5. Diagrammatic representation of the effects of **plasma concentrations** on **plasma clearances** in the human kidney.

Glucose is " reabsorbed " up to a transport maximum, Tm = 350 mg./min.

Diodone is " secreted " up to a Tm = 57 mg. (Iodine)/min.

Inulin is neither reabsorbed nor secreted in the tubules and the plasma clearance is little affected by the concentration in the plasma.

Some urea diffuses passively out of the tubules in amount about proportional to its concentration in the plasma. (After H. W. Smith.)

reabsorbed in the tubules so long as the amount presented to the tubule cells by the content of the glomerular filtrate does not exceed the transport maximum which in man is about 350 mg. per minute. Below this its plasma clearance is zero. If the plasma concentration exceeds a threshold value, about 0·18 per cent. for glucose, the excess which cannot be reabsorbed in the tubules appears in the urine, with measurable glucose clearances.

Proximal Tubular Activity

Micropipette samples of fluid from the glomerular capsule and from the termination of the proximal tubules have been obtained in frogs

and small mammals. After passing through the proximal tubule the glomerular fluid appears to have lost by reabsorption practically all its content of glucose and about four-fifths of its volume of water. The total osmolar concentration remains unchanged but creatinine which is not reabsorbed from the tubules is concentrated about five-fold because of the reduction in the volume of water. Since the osmolar concentration does not rise, the reabsorbed water must contain all the sodium and other ions in about the same isotonic concentration as in glomerular fluid, or rather higher, to compensate for increased concentration of little reabsorbed substances like creatinine and urea. Confirmation of these views can be drawn from studies of drastic diuresis in larger mammals. One form of diuresis, osmotic diuresis, can be produced by substances which are not reabsorbed in the tubules, because if more water were reabsorbed, their concentration would increase and the total osmolar concentration of the tubular fluid would not remain the same as that of the glomerular fluid ; the water, instead, flows on to become urine. Intravenous injection of large amounts of a powerful osmotic diuretic, such as a sugar called mannitol, may be followed by increase in the urine flow up to about one-third of the glomerular filtration rate. It is presumed that the flow down the later segments of the tubule is, then, so abnormally rapid that the composition of the outflowing urine will be practically the same as that of the fluid emerging from the proximal tubule, the distal tubule activity being swamped by the exceptional flood. Such urine has practically the same osmolar concentration as that of the plasma but it contains much more mannitol and much less sodium. Clearly, proportionately much more sodium than water is being reabsorbed, and since this sodium is being transferred from a weaker solution in the tubule to a more concentrated one in the plasma, it is inferred that reabsorption of sodium is an active process, known as a " sodium pump," which involves secretory work done by cells of the proximal tubule. The sodium ions transported by the sodium " pump " are accompanied by chloride ions, drawn passively out of the tubular fluid by the electrical gradient set up. In the proximal tubules, therefore, the secretory work necessary for reabsorbing the metabolically useful glucose and sodium chloride has the effect, indirectly, of also reabsorbing much of the equally useful water, leaving the unwanted creatinine, urea, and sulphates to be carried away in the urine.

The reabsorption of water in the proximal tubules is unaffected by the varying needs of the body and associated regulatory processes, but is decreased by osmotic diuretics. It has been estimated from indirect evidence that the average proportion of glomerular filtrate reabsorbed in the proximal tubule may be as much as 85 per cent. in the larger mammals and this is referred to as the " obligatory reabsorption " of water ; the fluid remaining corresponds in volume roughly to a maximal water diuresis. Further reabsorption of water in more distal tubular segments varies according to the requirements of the body and is referred to as " facultative reabsorption."

The Distal Tubular Complex

The loop of Henle, distal tubule and collecting duct form three anatomical segments which together convert the isosmotic effluent from the proximal tubule into urine. In overhydrated animals, the urine may equal this effluent in volume but will be much more dilute, nearly all the sodium chloride having been reabsorbed. In dehydrated animals, nearly all the water will have been reabsorbed but the urine that remains will be very concentrated, with a high content of sodium chloride, urea and other substances.

Micropipette samples of fluid from the first half of the distal tubule are hypotonic to blood, whether the urine is dilute as in water diuresis or concentrated as in antidiuresis. Here, or in the ascending limb of Henle's loop, therefore, must be a site of sodium chloride reabsorption. The distal tubule beyond, and also the collecting duct, must be practically water-tight in water diuresis. Since sodium chloride is still actively reabsorbed, a plentiful hypotonic urine is excreted. In antidiuresis, most of the water, as well as the sodium chloride, must somewhere be reabsorbed ; removal of the water-tightness of the collecting ducts and distal tubules seems to be the essential action of the pituitary antidiuretic hormone (A.D.H.) (Chapter 11).

When A.D.H. thus renders the collecting ducts permeable to water, the water comes under the influence of an osmotic gradient due to earlier reabsorption of sodium chloride in or before the distal tubules. Water is therefore passively reabsorbed, but this would continue only to the point of producing isosmotic urine were it not for the high concentration, much higher than that of systemic blood, in the tissue fluid and capillary blood in the medulla, progressively increasing as the tip of the papilla is approached. This marked increase in osmolar concentration of the inner sections of the kidney, compared with the isotonic cortex, has been demonstrated in rats by experiments on the freezing points of the tissues, microscopically observed, as well as by equilibration with salt solutions of tissue slices taken from various parts of the medulla. A high osmolar concentration has also been observed in blood withdrawn by micropipette from blood vessels in the papilla of the golden hamster. Thus the capillaries in the medulla take part in the countercurrent concentrating process described below.

The concentrating process in the kidney can, therefore, be explained as a primary process which raises the osmolar concentration of tissue fluid and all constituent structures of the innermost zone of the renal medulla, and a secondary process of water passively reabsorbed from collecting ducts as they pass through this zone, leaving a correspondingly concentrated urine to emerge from the papilla. This primary process of osmotic concentration in the papilla is attributed to the active reabsorption of sodium chloride in or before the distal tubule, some of the salt remaining in the tissue fluid and raising its concentration, instead of being carried away in the blood. That the concentration is graded, becoming intensified toward the papilla, is consistent with its being produced by the countercurrent mechanism described below.

The present short account of the formation of urine as due to two active reabsorptive processes, the proximal and distal sodium "pumps," provides the bare bones of the explanation for the concentrating and diluting processes. Much is also known about the ways in which urine with great variations in content of its individual solutes is produced to meet the varying needs of the body, but much of this knowledge is still embarrassed by disputed interpretations.

The "Hairpin Countercurrent" System. Suppose that a tube is divided along the middle by a septum which has the property of actively

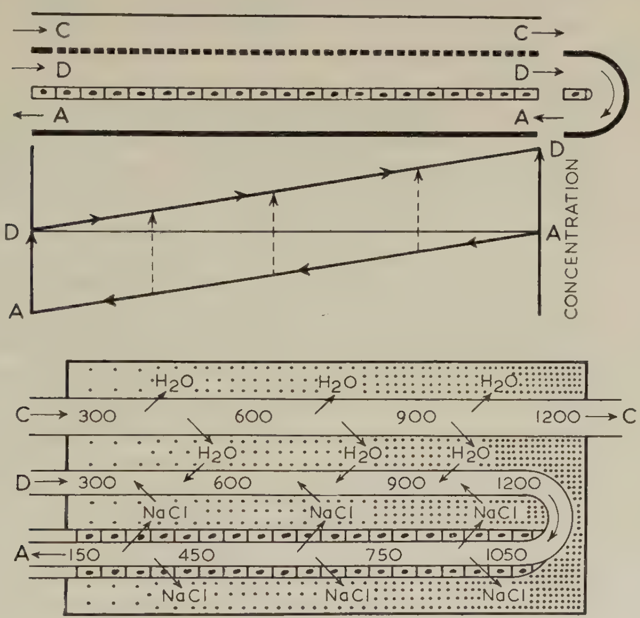


FIG. 9. 6. Diagrams illustrating the action of the **Hairpin Countercurrent System.**

At the top is an idealised system whose action in producing a great increase in concentration at the tip of the hairpin is described in the text. At the bottom is shown schematically the system as it exists in the kidney. A : ascending loop of Henle. D : descending loop of Henle. C : collecting duct. A is connected to C by way of the distal tubule, which actively re-absorbs NaCl, and allows water to pass from the tubular fluid to the blood so that the two become isotonic. The figures indicate the concentrations (milli-osmolar) of the fluid at various points within the loop of Henle and the collecting duct. The stippling indicates the corresponding increase in concentration of the interstitial fluid towards the tip of the loop.

transporting sodium chloride from the lower side A (Fig. 9. 6) to the upper side D. A solution of sodium chloride is supposed to flow from left to right along channel D, and from right to left along channel A : owing to the activity of the salt-transporting septum, the concentration of the solution in D will steadily rise, and that of the solution in A

will steadily fall, as is indicated in the middle part of the diagram, sodium chloride being transferred from one side to the other. But if, now, we join one end of these two channels together by a U-tube, the solution entering channel A will be the same as that leaving channel D : all the concentrations of the solution in A will be raised, as indicated by the vertical arrows, and will become nearly the same as those in channel D. We now have what is called a " hairpin countercurrent system," the reason for the name being clear from the diagram. The solution at the tip of the hairpin will be maintained at a much higher concentration than that of the solution entering and leaving it ; this excess concentration may be much larger than that which could be created by each element of the septum, if there were no countercurrent, and it might increase indefinitely with increase in the length of the whole hairpin system.

The solution leaving the hairpin has the same concentration as that entering it, so that by itself, it cannot yield a concentrated product. Suppose, however, that we now add channel C (Fig. 9. 6) which is separated from channel D by a membrane which is permeable to water, but not to any of the substances in solution. A solution flowing through channel C will lose water by osmosis into the hairpin system and thus become more concentrated : osmotic work will be done. Since the water removed from the solution in C is transferred to the solution in the hairpin system, it would be useless to connect the outflow from A directly to the inflow to C : this water must be removed from the system altogether, and not allowed merely to circulate within it. Further, the more water that is removed from C, the less concentrated is the solution within the hairpin, and the less effective it becomes in raising the concentration in C. For any considerable increase in concentration in C, the flow through C must be smaller than that round the hairpin.

Applying this idea to the kidney, we identify the hairpin system with the loop of Henle, the ascending loop being channel A, and the cells in its walls acting as the salt-transporting septum. These, however, do not transfer the sodium chloride directly to the fluid within the descending loop (channel D), but to the interstitial fluid, from which it may be supposed to diffuse passively through the thin walls of the descending loop as indicated in the lower part of Fig. 9. 6. Channel C may be identified with the collecting tubules whose walls must be supposed to be permeable to water, but relatively impermeable to all substances in solution. The loop of Henle, acting as a hairpin countercurrent system, will maintain a gradient of concentration in the interstitial fluid, increasing from the region of the convoluted tubules to the tip of the loop deep in the papilla. Such a gradient of concentration has, as already mentioned, been observed experimentally. As the urine flows down the collecting tubules it will lose water into the interstitial fluid and thus becomes more concentrated. This water must, of course, be carried away by the blood, or the whole process would soon come to an end. In the simplified system illustrated in Fig. 9. 6, the water is

shown entering the descending limb of Henle's loop, with the result that the fluid leaving the ascending limb becomes hypotonic. In the actual kidney, the process is somewhat different, owing to the presence of the blood capillaries which descend into the medulla, form loops, and return to the cortex. As in other low pressure capillaries, fluid will be drawn into the blood from the tissue spaces. The conditions for effective operation of a hairpin countercurrent system and its associated passive diffusing channel are thus present in the renal medulla, so that the urine in the collecting tubules can become considerably hypertonic. In water diuresis, however, when the antidiuretic hormone is absent, the walls of the distal and collecting tubules appear to become impermeable to water although sodium chloride is still reabsorbed by the " sodium pump "; a plentiful hypotonic urine is, therefore, excreted.

Acid-Base Regulation

Metabolism of food produces sulphate, phosphate, bicarbonate and hydrogen ions (Chapter 6, p. 190) which are eliminated in the urine. Most food, also, contains more potassium than sodium, so that there is relatively more potassium and less sodium in the urine than in the plasma.

Sulphate, phosphate and bicarbonate ions, like chloride ions, are drawn out of the fluid in the proximal tubules, in consequence of the active reabsorption of sodium ions ; but they move more slowly, and a much greater fraction is left behind and excreted in the urine. This fraction increases with the amount delivered to the tubules whether the increase be due to increase in glomerular filtration or in plasma concentration of the anions.

Potassium and hydrogen ions are also reabsorbed from the fluid in the proximal tubules but in the distal tubules they are transported in the opposite direction, from plasma to urine, by a process of " ion exchange." The sodium ions reabsorbed by the distal " sodium pump " are mainly accompanied by chloride ions but, in part, exchanged for potassium and hydrogen ions destined for the urine.

There is an inverse relation between the amounts of potassium and hydrogen ions excreted, for example, urine becomes alkaline when potassium salts are administered, and respiratory alkalosis produces increased potassium excretion. Moreover, acidosis due to inhalation of carbon dioxide mixtures decreases potassium excretion. This inverse relation is explained by supposing that potassium and hydrogen ions compete for a common secretory process in the tubule cells. Other such competitive secretory processes in the kidney are well known, one of them involving among other substances diodone, para-amino hippuric acid, and penicillin. Therapeutic use was made of this limiting process during war-time scarcity of penicillin when it was conserved in the body by administering carinamide which competes for the same renal secretory process, so that it delayed the excretion of penicillin.

There are two ways in which the acid end products of metabolism can be excreted in urine without undue lowering of *p*H which in man

never falls below 4·6. Firstly, the urine is well buffered mainly due to its phosphate content, the Na_2HPO_4 being converted to NaH_2PO_4. Other substances, such as bicarbonate in the alkaline range and urea in the acid range contribute. Secondly, persistent excretion of acid urine stimulates the formation of ammonia in the tubule cells and the neutralisation of acidic by ammonium ions.

Autoregulation of Renal Blood Flow

The blood flow through the kidney is very high, about 4 ml. per minute per gram of tissue. This compares with about 0·5 ml. for heart muscle and brain, 0·1 ml. for liver and 0·03 ml. per minute per gram of skeletal muscle. The blood volume in the kidney is also exceptionally

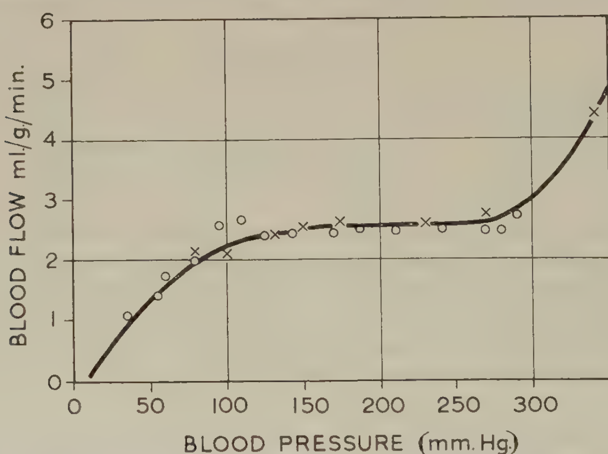


FIG. 9. 7. **Autoregulation of blood flow** between arterial pressures of about 100 and 280 mm. Hg in a kidney pump-perfused with blood from the carotid artery of the same dog.

Values obtained by continuous increase in pressure (circles) or by sudden increases in pressure (crosses) lie on the same curve, showing that the plain muscle of renal arterioles, unlike many other forms of plain muscle, does not contract after sudden stretch. (K. Thurau, K. Kramer and H. Brechtelsbauer, 1959.)

high, about 24 ml. per 100 g. tissue compared with about 8 ml. per 100 g. of the entire body.

In most organs, the blood flow varies directly with arterial pressure. Unless the rise of pressure is due to vaso-constriction, doubling the arterial pressure will rather more than double the blood flow because the vessels dilate with pressure. In the kidney, however, the blood flow rises very little with increase in arterial pressure so long as the pressure is within the physiological range of 80 to 180 mm. Hg. This control of blood flow is found in denervated and even in isolated perfused kidneys and is, therefore, called autoregulation.

Equally remarkable is the smallness of the rise in glomerular filtration rate when arterial pressure is raised, for example, by the carotid sinus

reflex in unanæsthetised animals. This again is an autoregulatory affair and sheds important light on the site of the varying resistance controlling the blood flow which must, therefore, reside in the pre-glomerular blood vessels.

Vasoconstriction of, or near, the vasa afferentia seems the obvious explanation of autoregulation, especially since the phenomenon may persist when kidneys are perfused with cell-free blood substitutes which cannot alter their viscosity, at any given temperature. Nevertheless, the hæmatocrit of the blood in the kidney at normal blood flows is barely one-half that in the systemic circulation and this must mean that the cells pass through the kidney, probably through shunts, more rapidly than does the plasma. Some experiments indicating that increase in pressure increases this division into cell-rich and cell-poor fractions of renal blood suggest that autoregulation may, in part, be due to an effective increase in the viscosity of blood with rise in arterial pressure.

The participation of blood in the increasing osmotic pressure within the medulla, as the papilla is approached, has been mentioned. It is clear from our discussion of the countercurrent concentrating process that an enormous blood-flow such as 4 ml. per minute per gram of tissue might swamp the process. It has been found, however, that the medullary blood flow is only about 1 per cent. of that of the cortex and, therefore, of the same order as the urine flow. Medullary flow is not subjected to autoregulation, as is the cortical flow and the increased medullary blood flow at high arterial pressure may, in part, account for pressure diuresis, since the increased blood flow limits the osmotic concentration in the tissues, and therewith, the withdrawal of water from the collecting ducts.

The Control of Renal Secretion

Under normal conditions, the kidney does not contribute to the variation of peripheral vascular resistance which maintains steady systemic arterial pressure. The sympathetic nerves which supply the organ only come into action when the arterial pressure falls below 60–80 mm. Hg when profound vasoconstriction and fall in glomerular filtration rate occur. This may be augmented by the effects of circulating adrenaline and noradrenaline under such conditions.

The effects on the secretion of urine of stimulation and section of various nervous structures may all be interpreted in terms of the changes induced either on the general arterial pressure, or on the calibre of the renal blood vessels, or on both. There is at present no sufficient reason for suspecting a direct nervous influence on the secretory mechanism proper, except for the anatomical fact that nerve fibres do supply the tubule cells. This is illustrated by the following experiment. If one kidney of a dog be completely denervated and the dog allowed to recover from the anæsthetic, the urine can be collected separately from each kidney through exteriorised ureters. The urine coming from the denervated organ is indistinguishable from that coming from its innervated fellow, both as regards rate of flow and composition. The

increase of urine flow due to administration of water (" water diuresis ") and its inhibition by exercise or stimulation of the skin are equal in both kidneys.

The degree to which circulating adrenaline and noradrenaline normally contribute to maintenance of the tone of renal blood vessels is uncertain. Larger concentrations, due to injection, produce vasoconstriction which, if slight, affects mainly the vasa efferentia and may

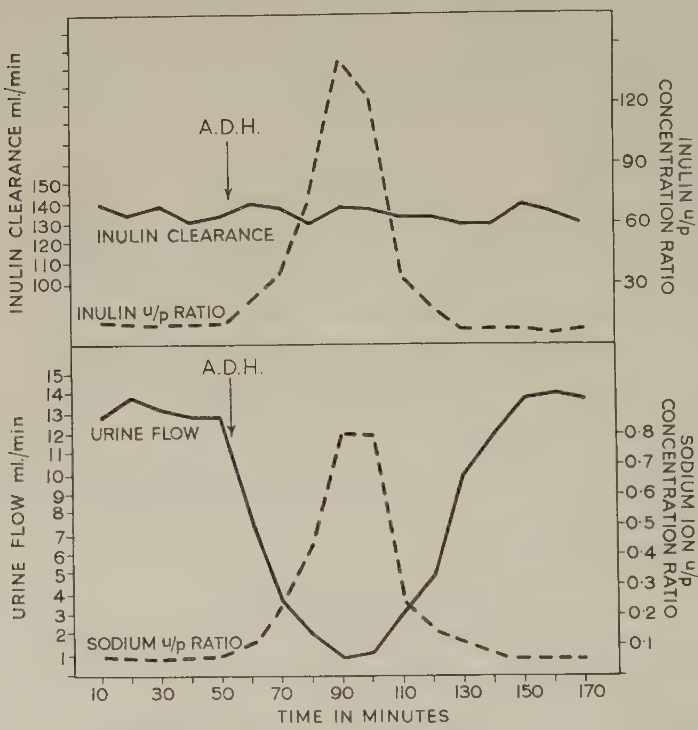


Fig. 9. 8 Intravenous injection of 40 milliunits of antidiuretic hormone produces in a well hydrated man (1) antidiuretic action with no change in inulin clearance, *i.e.* glomerular filtration rate (scale on the left), and (2) increase in urine concentration of sodium and inulin (scales on the right in terms of urine/plasma concentration ratio). Plasma inulin concentration 25 mg./100 ml. was kept constant by intravenous infusion of 1 g./hr. ; plasma sodium concentration : 325 mg./100 ml. (S. E. Dicker.)

raise glomerular filtration rate, but if profound, leads to fall in filtration and even anuria.

The antidiuretic hormone of the pituitary gland, as mentioned above, increases the permeability of the distal tubular complex and so increases the water reabsorption. Secretion of the antidiuretic hormone ceases if the body is hydrated by fairly rapid drinking of two or more litres of water with consequent reduction of the concentration of the body fluids by 3 to 5 per cent. A water diuresis of very dilute

urine ensues. Injection of hypertonic saline into the carotid artery stimulates secretion of the antidiuretic hormone with consequent oliguria and a hypertonic urine. The control of A.D.H. secretion is normally through osmo-receptors in the hypothalamus, but secretion may also be evoked by pain, fright or by administration of nicotine or acetylcholine. Ethyl alcohol is the only substance, other than water, known to inhibit A.D.H. secretion but it can also be inhibited by hypnotic suggestion of drinking water and it is reduced or absent in the disease diabetes insipidus.

Aldosterone and to a less extent desoxycorticosterone from the adrenal cortex have a profound effect on renal function by promoting reabsorption of sodium and excretion of potassium. Aldosterone inhibitors, *e.g.* spirolactones, therefore increase sodium output and consequently act as diuretics.

Any substance which cannot leave the proximal tubules as rapidly as water does will reduce the reabsorption of water and act as an osmotic diuretic ; for example mannitol, as already described, urea, sulphates, or an excess of glucose. The diuretics most widely used in medicine are, however, neither water, alcohol, aldosterone inhibitors nor osmotic diuretics but substances, such as chlorothiazide or organic mercury compounds, *e.g.* mersalyl, which appear to act by inhibiting chloride or sodium chloride reabsorption. Their diuretic action is sometimes reinforced by combining them with one of the purine diuretics, aminophylline or caffeine, which may increase glomerular filtration by dilating the vasa afferentia and may also reduce tubular reabsorption.

MICTURITION

Urine is secreted by the kidneys continuously, but removed from the body only periodically. Meanwhile it collects and may remain for some hours in the bladder, which acts as a reservoir.

The urine passes from the kidneys to the bladder through tubes, the ureters, the walls of which contain plain muscle, and encourage the downward flow by peristaltic contractions. These contractions travel down the tube at about 2 to 3 cm. per second, and are repeated from about one to four times a minute. Consequently the urine enters the bladder in a series of squirts, as can be observed in man by looking through a cystoscope, a tubular instrument inserted through the urethra, which illuminates and renders visible the lining wall of the bladder. If, for any reason, the pressure inside the ureter is raised, its contractions become fiercer and more spasmodic in nature, a change which is readily observed in the isolated kidney of the dog. Such spasmodic contraction of plain muscle produces intense pain, and is illustrated in human disease by renal colic, which is an attack of severe pain and other symptoms due, for example, to passage of small stones from the kidney, blocking and distending the ureter and so evoking spasm of its wall.

Evacuation of the bladder is effected by contraction of the plain muscle in its walls, known as the detrusor muscle. The urethra is

guarded by two sphincters—the external being of striated and the internal of unstriated muscle. If the external voluntary sphincter be held open by a catheter, the vesical contents are retained by the internal involuntary sphincter ; even in such circumstances the subject can pass urine voluntarily by initiating a contraction of the detrusor muscle which induces relaxation of the sphincter. As the bladder fills, the intravesical pressure in man rises to about 5 cm. water at a volume of 200 ml. ; the pressure is maintained at about this value until the volume of urine reaches 400 ml., after which further filling brings about a

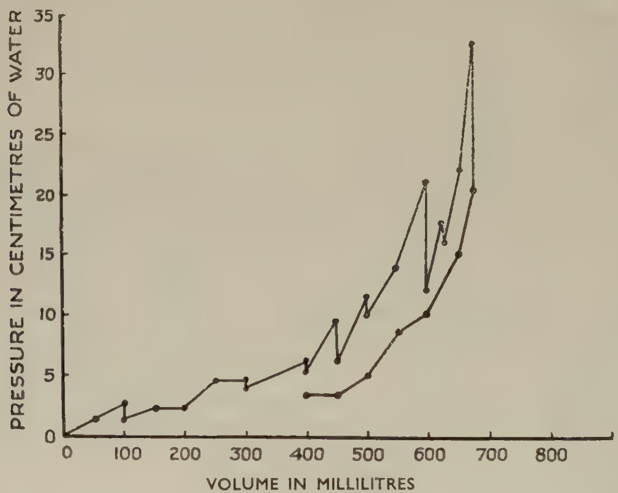


Fig. 9. 9. Changes in the **pressure in the bladder** of a man during filling and emptying.

Upper Curve. Water was slowly run into the bladder, and the pressure was observed after the addition of each 50 ml. The inflow was stopped at intervals, so as to allow time for the pressure to approximate to its final value, as shown by the short vertical lines.

Lower Curve. After 700 ml. had been run in, the bladder was allowed to empty, 50 ml. at a time. The pressures were all lower, indicating that the true equilibrium values had probably not been reached during either filling or emptying.

The pressure is not strictly constant over any range of volumes, but does not vary much between 100 ml. and 400 ml. (Denny-Brown and Robertson.)

steeper rise in pressure, 20 to 30 cm. water being reached when the volume is about 500 to 600 ml. (Fig. 9. 9).

The property of maintaining a constant pressure within the bladder through this range of volume was at one time attributed to an active relaxation of the muscle while the organ was being filled. The constancy is due, however, largely to the purely physical relation whereby, when a hollow organ is distended, and the length and tension of the muscle fibres in its wall are correspondingly increased, the pressure of liquid inside may hardly rise at all because the curvature of the wall becomes progressively less. A similar effect reduces the rise in pressure inside a toy balloon when it is blown up.

The bladder, like the alimentary canal, has a double autonomic innervation. Stimulation of the parasympathetic (pelvic) nerves induces contraction of the bladder but relaxation of the internal sphincter; the parasympathetic system contains the most important pathways for reflexes concerned in micturition including both afferent and efferent elements in the arcs. Stimulation of the sympathetic (pre-sacral or hypogastric) nerves causes contraction of the internal sphincter; relaxation of the bladder wall may follow intravenous injection of adrenaline in man.

The desire for micturition is set up when the volume of urine in the bladder reaches 200 to 300 ml., and a certain tension in the muscular wall is reached. When the tone of the bladder is increased, *e.g.* by cold or conditions of emotional strain, this critical tension may be reached with a relatively small volume of urine in the bladder. Up to a certain point, the sensation from a full bladder can be suppressed from a conscious level in the cerebral cortex, and, on the other hand, it is possible, by introspection, to become aware of small quantities of urine in the bladder.

Voluntary micturition is brought about by impulses passing from the cerebral cortex by way of the spinal cord and parasympathetic nerves, to the bladder. The detrusor muscle contracts strongly, raising the pressure in the bladder to 100 cm. water, and simultaneously a reciprocal relaxation of the internal sphincter occurs. Voluntary effort to restrain micturition may considerably reduce the pressure within the bladder. Thus we have an example of involuntary muscle under the control of the will. Once micturition has begun, certain reflexes play a part in its completion. (1) Stretching of the bladder wall brings about reflex contraction of the bladder and this reflex can be abolished by cutting both the pelvic nerves, transecting the spinal cord, or cocainising the interior of the bladder. It is unaffected by division of the hypogastric nerves. The reflex arc concerned is along the pelvic nerves to a centre in the hind-brain and back again along the pelvic nerves. (2) The flow of water through the posterior part of the urethra also brings about reflex contraction of the bladder, the reflex arc involving the centre in the hind-brain. Transection of the central nervous system only interferes with reflex micturition if the section is below a plane passing from the inferior colliculi dorsally, to the middle of the pons ventrally.

Diseases of the spinal cord involving the posterior columns may prevent the sensations from the bladder reaching consciousness, although micturition can still be carried out voluntarily. When the pyramidal tracts are interrupted, voluntary micturition is impossible, although the patient may be quite aware of a full bladder. When the voluntary control of micturition is impaired, retention of urine may result in over-distension of the bladder, and reflex passage of small quantities of urine at irregular intervals; this condition is known as "retention with overflow."

CHAPTER 10

REPRODUCTION

In all but the lowest mammals, the young spend the first part of their life—forty weeks in man—as parasites of the mother, protected by her body from the physical and chemical vicissitudes of the outside world and provided from her blood stream with an unfailing supply of food and oxygen. The *uterus*, the thick-walled hollow organ in which this early development takes place, receives the egg from one of two narrow symmetrical ducts, the *Fallopian tubes*, which open into its upper part. Their free outer ends are adapted to receive eggs from the *ovaries* when these organs discharge them. The narrow lower part of the uterus—the *cervix*—projects into and inverts the dome of the *vagina* and is traversed by a narrow canal through which uterine and vaginal cavities communicate. The cells lining this canal secrete a mucus through which pass the *spermatozoa*, which, produced in the *testis* of the male, are ejaculated from the *penis* into the upper end of the vagina during coitus. Should coitus coincide sufficiently closely with ovulation, one of the highly motile spermatozoa may reach the egg in the outer part of the Fallopian tube and penetrate its envelope so as to fertilise it. Following fertilisation, the egg subdivides and, passing down the Fallopian tube, enters the uterine cavity where it later becomes embedded in the uterine mucosa, or *endometrium*. The period of *pregnancy* then follows.

Within a short time, the embryo develops a blood circulatory system which comes into intimate relation but does not mix with that of the mother. This close association of the maternal and embryonic circulatory systems takes place in a fœtal structure, called the *placenta*, through which food, oxygen and other substances can diffuse from the mother to the fœtus. Throughout pregnancy, the fœtus and its associated placenta and membranes grow, the uterus enlarging to accommodate its expanding contents. The process of birth, or *parturition*, at the end of pregnancy is brought about by contractions of the muscular wall of the uterus, following dilatation of the cervical canal, which permits the expulsion of the fœtus. Following parturition, the young receive for a time the whole or part of their nourishment from milk secreted by the mother's *mammary glands*.

After a variable period of growth, in man eleven to fourteen years, the phase of puberty begins. During this, the ovaries of the female commence to discharge ova and the testes of the male to produce spermatozoa. Other important changes, psychological and physical, occur and **secondary sexual characters** appear or are emphasised. In girls, enlargement of the breasts, deposition of subcutaneous fat and pelvic skeletal changes determine the typical feminine contour. Pubic

hair, typically with an abrupt horizontal border in the female, and axillary hair appear in both sexes. In boys, the beard begins to grow and the voice breaks. These changes result from the presence of the **sex hormones**—androgens in the male and œstrogens in the female—secreted by the testes and ovaries respectively, under the influence of the gonadotrophic hormones from the anterior pituitary. Although procreation is possible at this stage of development, full maturity in man is not considered to occur until five or ten years later.

The **primary reproductive organs,** or **gonads,** are the testes of the male and the ovaries of the female. The accessory reproductive organs of the male include the *epididymides*, the *vasa deferentia*, the *seminal vesicles*, the *prostate gland*, *Cowper's glands* and the *penis*. Those of the female include the *Fallopian tubes*, the *uterus*, the *vagina*, *Bartholin's glands*, the *clitoris* and the *mammary glands*. While mammals are similar in their primary and accessory organs, they are widely diverse in their secondary sexual characters.

Spermatogenesis

The testis consists largely of *seminiferous tubules* (Fig. 10. 1) which, in man, have an aggregate length of about one thousand feet. Spermatozoa are produced within the tubules and the ciliated *vasa efferentia*

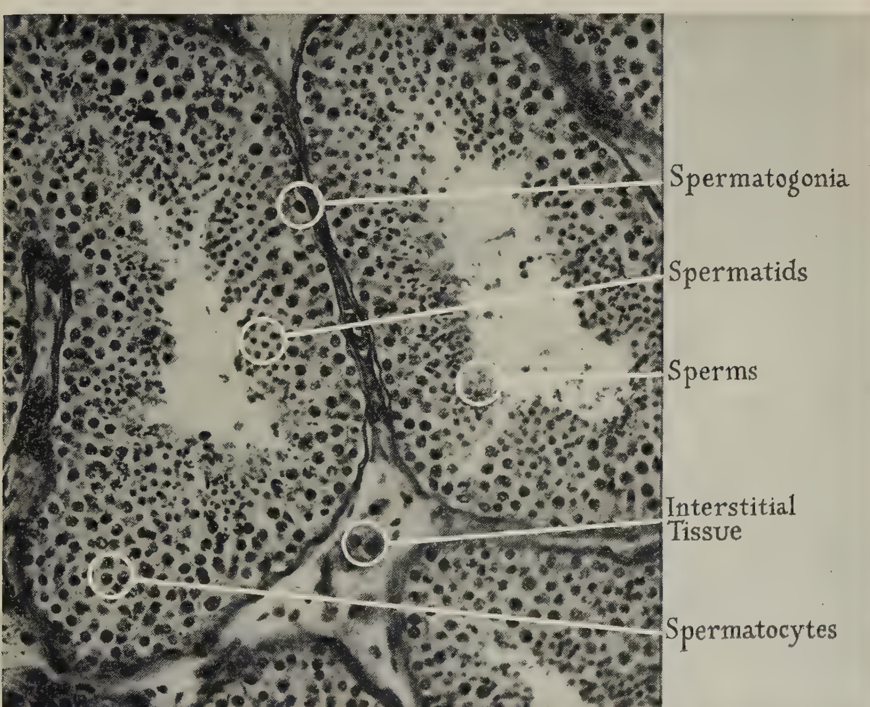


FIG. 10. 1. Microscopic section of Human Testis, showing seminiferous and interstitial tissue. (From a preparation by Mr. K. Richardson.)

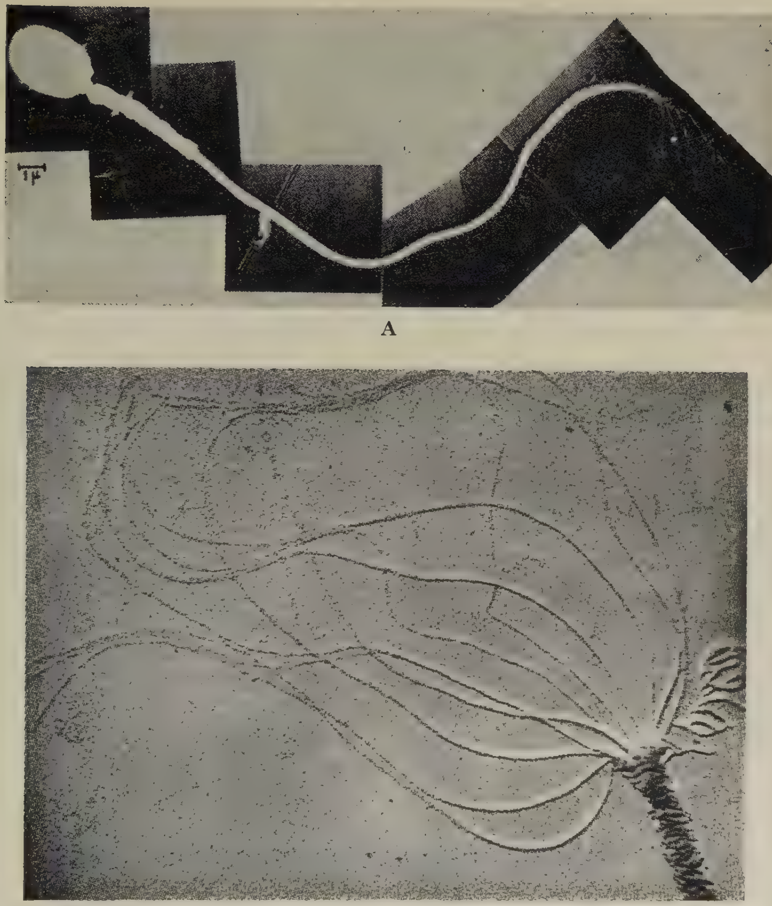
pass them on the *epididymis*, a tube about fifteen feet long where they are stored. They are also stored in the ampulla of the vas deferens but not in the seminal vesicles.

The tubules are lined by *germinal epithelium*, derived from the repeated division of the primordial germ cells. These give rise to *spermatogonia* and *Sertoli cells* which are the only forms found before puberty. At puberty, however, differentiation of the germ cells begins. Under the influence of pituitary gonadotrophic hormones, the spermatogonia divide to form *primary spermatocytes*, each of which undergoes only two further divisions (maturation divisions) and therefore gives rise to only four further descendants. The first of these divisions is the *reduction division*, or *meiosis*, in which the number of chromosomes, characteristic for the species, is halved. In man, the chromosome number is 46 (Fig. 10. 10, p. 300), so that each *secondary spermatocyte* contains 23 chromosomes. The formation of a tail now begins and the cell is called a *spermatid*. It is generally believed that the spermatids become attached to the Sertoli cells (also called *nurse cells*) in order to develop further into spermatozoa. The details of this transformation are intricate ; the gross changes consist in great elongation of the tail and gradual loss of the protoplasm surrounding the nucleus.

Mature spermatozoa of different species show enormous morphological variation. The human spermatozoon (Fig. 10. 2A) consists of a head 4·6 μ long and 2·6 μ wide and of a caudal appendage 35 μ to 45 μ long. The head contains the nucleus at its base. The caudal appendage includes the midpiece, tail and terminal filament. The midpiece is about the same length as the head and is traversed by the axial filament, covered by a sheath. Around this is coiled the helical filament of five or six turns contained within a cylindrical sheath. The tail, 30 μ to 40 μ long, consists of the axial filament covered by a thin sheath. The axial filament contains some 11 fibrils and continues as the terminal filament (Fig. 10. 2B) uncovered by the sheath. The latter is composed of helically coiled fibrils

The fibrillar structure of the midpiece and tail is undoubtedly the basis of the motile mechanism of the spermatozoon. Contractions of the fibrils cause a wave-like bending of the tail from side to side, proceeding downwards from the midpiece, and propelling the spermatozoon in a corkscrew path at a rate which, in man, is about 3 mm. per minute, but which varies greatly in different species.

The spermatozoa leave the testis before maturation has been completed and the final stages are usually passed in the epididymis. It is here that the spermatozoa acquire motility. Unless ejaculation takes place, the accumulated spermatozoa within the epididymis eventually die and are dealt with as foreign bodies (Chapter 21, p. 610), being engulfed by macrophages and digested. If the vas deferens is blocked, or ligated, spermatogenesis continues and this mechanism (to which the name of " spermatophagia " has been given) provides for their removal. If the vasa efferentia are ligated, the back pressure effect on the testis is so great that destruction of the germinal epithelium usually



A

B

FIG. 10. 2. The Human Spermatozoon.

A. An intact spermatozoon reconstructed from 7 separate electron micrographs.

B. The 11 fibrils of the axial filament emerging from the sheath as the terminal filament. The sheath consists of helically coiled fibrils. (Bayle, *Proc. Soc. Study of Fertility.*)

follows. In the guinea pig, spermatozoa may remain motile for two months after ligation of the vas deferens but they lose their fertilising capacity after one month. In the rabbit, functional spermatozoa have been found as much as a year after ligation of the vas deferens.

Repeated ejaculation rapidly reduces the number of spermatozoa in the ejaculate. Recovery to the normal number may require up to five to seven days.

Spermatogenesis in most lower animals is seasonal. In domesticated and semi-domesticated animals and in man and some other primates it is continuous. It is dependent on the secretion of gonadotrophic

hormones by the anterior pituitary gland. In the absence of normal pituitary development, spermatogenesis never begins and if the pituitary gland is removed from an adult animal, spermatogenesis comes to an end. Implantation of fresh pituitary glands or the injection of suitable extracts in hypophysectomised animals leads to recovery of spermatogenesis or, in young animals, to precocious spermatogenesis. In those animals which exhibit seasonal activity, climatic factors, such as light and heat, mediated by the central nervous system and pituitary gland, are responsible.

The temperature of the scrotum is from 1 to 8° C. lower than that of the abdomen. The germinal epithelium of the seminiferous tubules degenerates if the testes are removed to a warmer environment (for example, inside the abdomen) or if the scrotum is experimentally insulated against heat loss. For a few months, recovery is possible but the change ultimately becomes irreversible. Heating the scrotum beyond body temperature quickly injures the germinal epithelium and prolonged pyrexia may be followed by temporary sterility. Only if they are transplanted into exposed situations do testicular grafts in animals show spermatogenesis. The dartos muscle of the scrotum and the cremaster muscles, by relaxing with heat and contracting with cold, help the thermo-regulatory mechanism. Failure of testicular descent (cryptorchidism) is a cause of sterility in the human. Unless the testicles can be brought into the scrotum well before the onset of puberty, failure of spermatogenesis is the result and cannot be corrected by later surgery.

Of **environmental conditions** influencing spermatozoa, the most important is hydrogen ion concentration, the optimum pH for preservation being a little over 7·0. Other ions may exert an important but less fully investigated effect. As the pH rises up to 8·5, sperm motility increases but this exhausts the small available supply of energy. It is often said that epididymal spermatozoa are quiescent but this is not true ; when, on operation on men with blockage of the tail of the epididymis (caused by gonorrhoea), the distended head of the epididymis is incised and a drop of fluid from it is examined under the microscope, highly active spermatozoa are often found. Prostatic secretion (pH 7–8) increases, vaginal secretion (acid) diminishes and cervical secretion (alkaline) increases motility.

Body temperature is optimal for the motility, a lower temperature for the preservation of spermatozoa. The significance of the natural secretions which they meet is not known for certain, though there is little doubt that these do have some effect on the fertilising capacity of the sperms. It has clearly been demonstrated that spermatozoa must pass a certain minimal time within the Fallopian tubes, during which they undergo an effect called *capacitation*, before they become capable of fertilising an egg. However, experimental studies have shown that various other organs, apart from the Fallopian tubes, will also serve as vehicles for this capacitation effect. Although viability of spermatozoa within the epididymis is well supported, no known pro-

tective substance has been extracted from epididymal secretions. Within the female reproductive tract, spermatozoa are rapidly killed in the vagina because of its high acidity but they may survive well in the cervical secretions around the time of ovulation. They survive best, however, in the Fallopian tubes from which, in the human female, live spermatozoa have been obtained up to seven days after the last occasion of coitus. Within the cervix, survival seldom extends beyond seventy-two hours and is, of course, often very much less than this. There is little doubt that the ability to fertilise is lost before motility ceases.

There is little doubt that spermatozoa reach the Fallopian tubes in comparatively short time. In some species, this is due to mass entry of semen into the uterus and its propulsion through the tubes by contraction and relaxation of their walls. This, for example, is known to occur in the mare, since certain of the chemical constituents of seminal fluid have been detected in the Fallopian tubes shortly after copulation. Although the occurrence of " cervical insuck " has been postulated in the human female, it is unlikely that such a mechanism exists and more probable that sperm entry into the cervical secretions depends largely on their own unaided efforts and that passage through the uterus is similarly unassisted. Passage through the Fallopian tubes is quite likely aided by the activity of the cilia lining the tube and by the muscular contractions of the tubal wall. Although the cilia lash in a direction opposite to that which the sperms have to travel, and can be shown experimentally to cause the passage of particles down the tube from the fimbrial end towards the uterus, reverse currents also exist and these may assist transport of spermatozoa up the tube. The normal site of fertilisation of the egg is in the ampulla of the tube.

The Ovum and Ovulation

The precise source of the *primordial egg cells* is uncertain but the modern view is that they are derived from specialised cells which, during development of the embryo, migrate from the yolk-sac endoderm into the germinal ridge and therein cause the formation of *primordial follicles*, consisting of the ovum itself (oöcyte) surrounded by follicular epithelial cells. All the ovarian (or *Graafian*) follicles are formed during embryonic life and thereafter no further formation of oöcytes occurs. This view is in opposition to an earlier view that formation of new oöcytes (oöneogenesis) occurred from the germinal epithelium, the cell layer covering the ovary, during postnatal life. Accurate experimental observations have failed to support this hypothesis but are in favour of the first-mentioned one. The number of primordial follicles is greatest at about the time of birth. Thereafter disintegration of the follicles—called *atresia*—occurs on a large scale before puberty but later on a lesser scale. In the human, both ovaries contain perhaps 400,000 primordial follicles at birth, of which only some 400 are destined to take part in ovulation, the remainder undergoing atresia.

Enlargement of follicles occurs continuously due to proliferation of the cells surrounding the ovum, which differentiate into *granulosa cells*

and a capsule of two layers, called the *theca externa* and the *theca interna.* The ovum itself is a large cell with a well defined limiting membrane (the *zona pellucida*) and a nucleus with a distinct nucleolus. Until puberty, follicular maturation does not occur, the larger follicles undergoing the atretic process mentioned above. At puberty and afterwards, under the influence of gonadotrophic hormones from the anterior

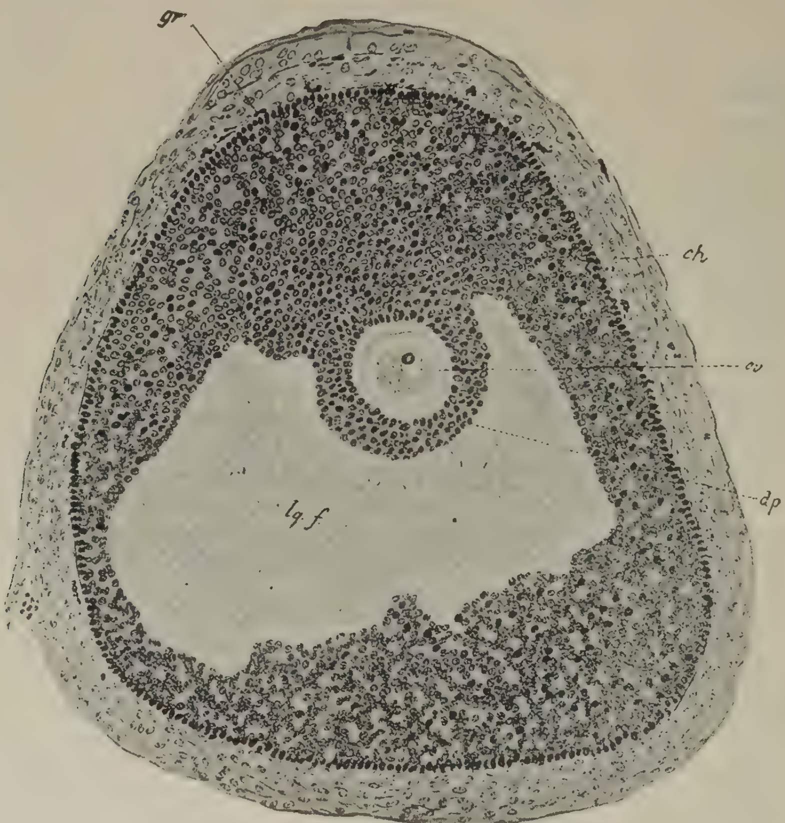


FIG. 10. 3. Graafian Follicle of mammalian ovary. (Prenant and Bouin.) *Ov.* Ovum. *dp.* Cumulus. *ch.* Theca. *gr.* Membrana granulosa. *lq. f.* Liquor folliculi.

pituitary gland, full follicular maturation occurs. In this, due to secretion of a fluid, *liquor folliculi*, by the theca cells, the follicle becomes a cyst, in which the ovum is found embedded in a cone of follicle cells, the *cumulus* (Fig. 10. 3).

Ovulation. After puberty, although many follicles continue to become atretic, others reach maturity (Fig. 10. 4) and ovulate. Ovulation begins with a second and more rapid phase of follicular enlargement, during which the ovum becomes detached from the wall of the follicle,

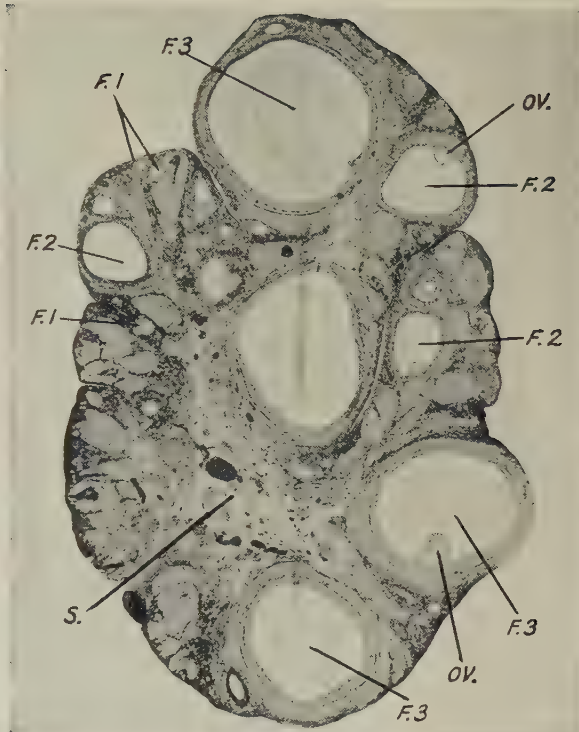


FIG. 10. 4. Section through the **Ovary** of a **Ferret just before Ovulation.**

F. 1, very immature follicles ; *F.* 2, developing follicles ; *F.* 3, mature follicles ; *OV*, ovum ; *S*, stroma tissue. (From a photograph supplied by Dr. Parkes.)

FIG. 10. 5. Section through the **Ovary** of a **Mouse during Di-œstrus**

F. 1, immature follicles ; *C.L.*, corpora lutea ; *C.*, capsule of ovary. (From a photograph supplied by Dr. Parkes.)

retaining follicular cells around it—the *corona radiata.* It undergoes reduction division and extrudes the first polar body. The follicle has now a very thin outer wall and projects from the ovary. In man it is about the size of a pea. At its thinnest point, the follicular wall bursts or is dissolved and the ovum, leaving the ovary, is picked up by the fimbriated end of the Fallopian tube which it enters. The rupture in the ovary heals and the follicular cells proliferate to form the *corpus luteum.* Hypertrophy of the granulosa cells begins and in a few days they become very large and contain a yellow lipoid material, lutein. Similar changes occur in the cells of the theca interna. The luteal cells tend to fill up the follicle (Fig. 10. 5) but a central cavity persists in the human corpus luteum for some time, being filled with liquor folliculi and a small amount of blood. Regression of the corpus luteum, which is much delayed during pregnancy, begins with fatty infiltration of the cells which are ultimately displaced by a hyaline substance, forming a corpus albicans and finally shrinking to a scar. The principal functions of the corpus luteum and the timing of ovulation in relation to the other events of the reproductive cycle are dealt with on pp. 291 *et seq.*

In many animals, including man, in whom it occurs monthly, ovulation takes place without any external stimulus. In others, like the rabbit, cat and ferret, it is secondary to copulation which it follows in about ten hours, the ripe follicles otherwise becoming atretic. If the pituitary is removed within one hour of copulation in these animals, ovulation fails. The remaining nine hours must be occupied by the final processes of ovulation in the ovary for hypophysectomy in this interval does not arrest it. Normal ovulation, therefore, whether spontaneous or secondary to coitus, is probably due to the action of the pituitary gonadotrophins (p. 294).

After ovulation, the ovum remains capable of being fertilised for a short time only. In most species this is probably less than twelve hours.

The Endocrine Functions of the Testis and Ovary

The primary reproductive organs, in addition to producing sperm or ova, respectively, also secrete hormones—the " sex hormones "—which are responsible for the changes which occur at puberty in the accessory reproductive organs, and for the development of the secondary sexual characters. The male hormones are known collectively as " androgens " and the female hormones as " œstrogens." These internal sections are under the control of the gonadotrophic hormones, since after removal of the pituitary, the secretion of androgens and œstrogens ceases.

The substance *testosterone* has been isolated from testicular tissue, and *œstradiol* from the ovaries. These two substances are generally regarded as the principal male and female sex hormones respectively but it is now well known that many other, naturally-occurring compounds, all having the *cyclo*pentenophenanthrene nucleus of the steroids, possess androgenic and œstrogenic properties (Fig. 10. 6). In addition,

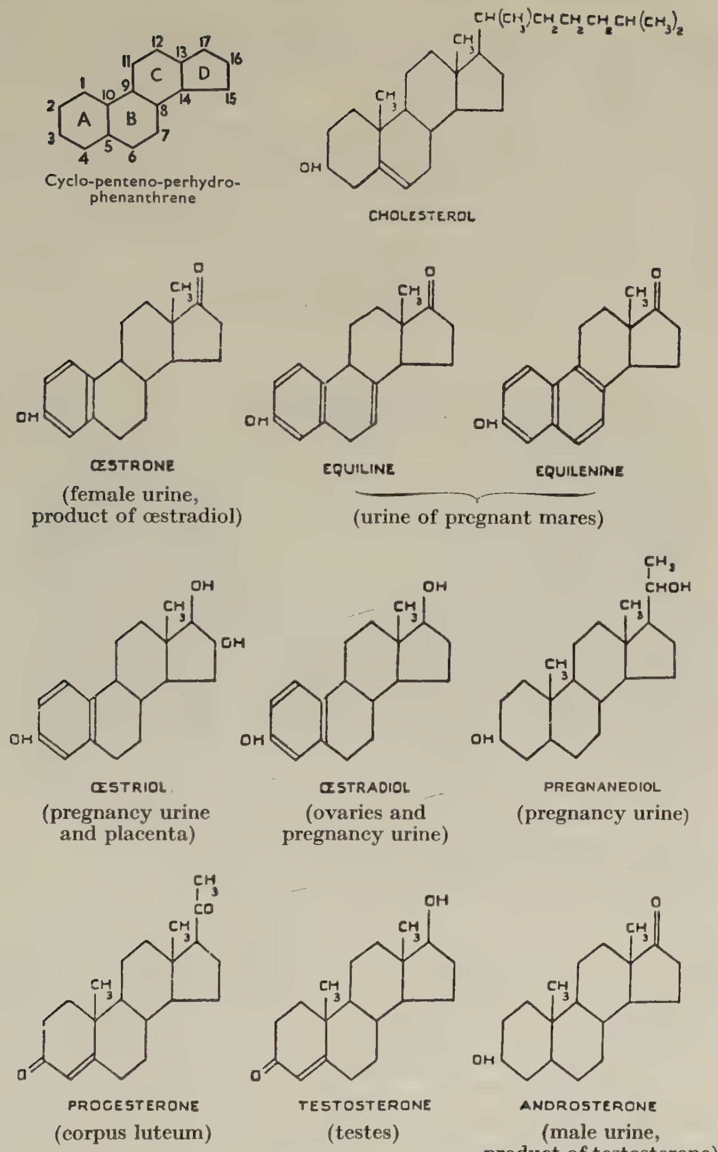


FIG. 10. 6. The chemical relationships of the sex hormones, and the organs and fluids from which they can be obtained.

a number of semi-synthetic androgen and œstrogen derivatives, as well as wholly synthetic non-steroidal œstrogens, have been prepared by the organic chemists. The principal sources of natural œstrogens are urine, especially that of the stallion and of pregnant women. They are excreted mainly in the form of glucuronides and sulphates ; as

such, they are water soluble and physiologically virtually inactive. It is not known for certain in what form they are secreted into and carried by the blood stream but it is probable that, to a greater or lesser extent, they are bound to plasma proteins.

The Endocrine Function of the Testis. If the testes are removed *before puberty*, the changes which usually occur at this time, in the accessory sex organs and the secondary sexual characters, do not appear. In man, the penis remains small and the epithelium of the seminal vesicles, prostate and Cowper's glands remain undeveloped. Secondary sexual characters fail to appear, the voice continues to be high-pitched and the facial and body hair remain scanty.

Castration *after puberty* causes rapid regression of active epithelia in

Fig. 10. 7. Effect of castration and testis hormone on epithelium of seminal vesicles. (Moore, Hughes and Gallagher.)

(1) Cells from normal animal, showing secretion granules.
(2) Cells from twenty-day castrate.
(3) Cells from twenty-day castrate treated with male hormone.

the accessory organs (Fig. 10. 7), but the more obviously permanent pubertal changes, such as enlargement of the penis and alterations in the larynx persist. Erection and copulation may continue for years in some, but not all, post-pubertal castrate men, though *libido*—the sex urge—is usually impaired to a greater or lesser extent. What is called " sex drive " in animals rapidly falls off after castration.

Many other effects are ascribed to castration, some of which are undoubtedly primary. In domestic animals, for example, the modifications of temperament and edibility are sufficiently important to make the practice widespread. In man, however, the sphere of the mind is immeasurably larger than in animals and we learn by common experience that the same bodily defects may render one man useless, yet make little difference to another. Furthermore, a person may suffer more pain and anxiety on account of a non-existent disease which he fears than does another suffering from the real thing. The stigmata associated with castration vary so much, and are so deeply rooted in tradition and

superstition, that where physiological and psychological effects can interact, it is impossible to disentangle them. A certain degree of "effeminacy," for want of a better word, is frequently discernible in prepubertal castrates.

Administration of androgens to castrates, according to whether they are pre- or post-pubertal, develops or restores the epithelia and the size of the accessory organs and, if previously absent, secondary sexual characters appear. Even in post-pubertal castrates, the penis may enlarge, erection be facilitated and the voice become deeper. Changes in personality, as well as in physique, occur ; the treated castrate assumes a more dominant, purposeful character, and at the same time, muscle mass and strength increase, due to the protein anabolic activity of androgens. There is in consequence a distinct sense of "well being" and a feeling of physical fitness.

Since the effects of removing the testes can be annulled by administering androgens, which are known to occur in testicular extracts, the testis must be the sole *effective* source of these substances ; in virtue of this side of its activity it is thus an important endocrine organ. Destruction of the spermatogenic cells of the testis (*e.g.* by heat, as in cryptorchidism, or by X-irradiation) does not abolish its endocrine activity, for which there is no doubt that the interstitial cells are responsible. A not uncommon human testicular disorder is *germinal aplasia*, in which the seminiferous tubules contain Sertoli cells but no germ cells. The interstitial cells are normal and such men, though sterile, show no defects of masculinisation.

Assay of Androgens. For the biological assay of androgens, the comb of the castrated cock (*capon*) has been most widely used. The androgens, in a suitable oily vehicle, are applied directly to the comb by inunction and its growth is measured under standard conditions. Other test objects have been the weight of the prostate and seminal vesicles in castrated male rodents, and the fructose content of seminal fluid. This fructose is a secretion of the seminal vesicles and the quantity secreted is under the control of androgens.

The Endocrine Functions of the Ovary and the Menstrual Cycle. Removal of the ovaries causes failure of development of the accessory sex organs—the uterus, vagina and mammary glands. Axillary and pubic hair fail to grow, the distribution of fat does not assume the typically feminine pattern and skeletal changes in the pelvis do not occur. After puberty, regression of the accessory organs takes place. These deficiencies can be prevented by the administration of œstrogen or, in animals, by the transplantation of ovarian tissue.

Although the ovary has been proved to be the principal source of œstrogens and these to be essential for reproductive activity in the female, it is not so easy to determine their cellular origin. In some animals (*e.g.*, the mouse) all the follicles may be destroyed by X-rays, owing to the peculiar susceptibility of rapidly growing tissues to this agent, and the secretory activity of the ovary remains unimpaired. In other animals, the internal secretion of the ovary fails under this treatment, and we shall see that there is commonly a high degree of

correlation between follicular and œstrogenic activity. It is probable that the follicle cells (the theca interna cells are, it is thought, most likely the ones concerned) secrete œstrogen, but other actual or potential sources of œstrogen are interstitial cells of the ovary and the adrenal cortex. Evidence from human beings, derived from studies of patients with carcinoma of the breast who have undergone the operation of oöphorectomy, has shown clearly that although the excretion of œstrogens may fall after the operation, within a comparatively short time the level rises and this is due to increased œstrogen production by the adrenal cortex. Adrenalectomy causes a further fall in œstrogen output but not uncommonly even this fails to abolish completely the excretion of œstrogens. It is supposed that some adrenal tissue may remain and may be stimulated to increased activity by pituitary gonadotrophins, once the adrenals themselves and the ovaries have been ablated.

The most striking event after oöphorectomy in the mature female is, however, the disappearance of the *sexual cycle*. In women this takes the form of **menstruation**, the loss from the uterus about every twenty-eight days for a period of three to five days of fluid containing a high proportion of altered blood.

In animals, the cycle is marked not so much by visible structural changes as by recurring changes in behaviour indicating increased sexual receptivity, when they are said to be " on heat " or in *œstrus* (only at this time will they receive the male). The cycle in animals is therefore known as the *œstrous* cycle.

In spite of the different nature of the events by which the cycle is marked externally, it has been found in both man and animals to be based on similar events in the ovaries. At œstrus, animals have ripe Graafian follicles and their secretion of œstrogens is at a maximum, as evidenced by hypertrophy of the uterus and vagina and by changes in behaviour, all of which can be reproduced by giving œstrogen to a spayed animal. Changes in the vaginal epithelium are useful for following the cycle in rodents, as shown in Fig. 10. 8. Cornification of the vagina of spayed rats, detected by means of the vaginal smear, is an important biological test for œstrogen, comparable to the capon's comb test for androgen. Ovulation, whether spontaneous or depending on coitus, occurs during œstrus.

There is some evidence that small amounts of progesterone (see below, p. 291) are secreted prior to ovulation and act synergistically with œstrogen in causing œstrous behaviour. In spayed animals, the administration of small amounts of progesterone permits the induction of œstrus with much smaller quantities of œstrogen than would be the case with œstrogen alone.

The maximum ripening of follicles in the human female, though associated with a peak of œstrogen secretion, passes relatively uneventfully for the individual about twelve to fourteen days after the beginning of the last menstrual flow. The external event of the human cycle—menstruation—occurs, therefore, when œstrogen secretion is low and to understand its significance we may examine the changes which sex

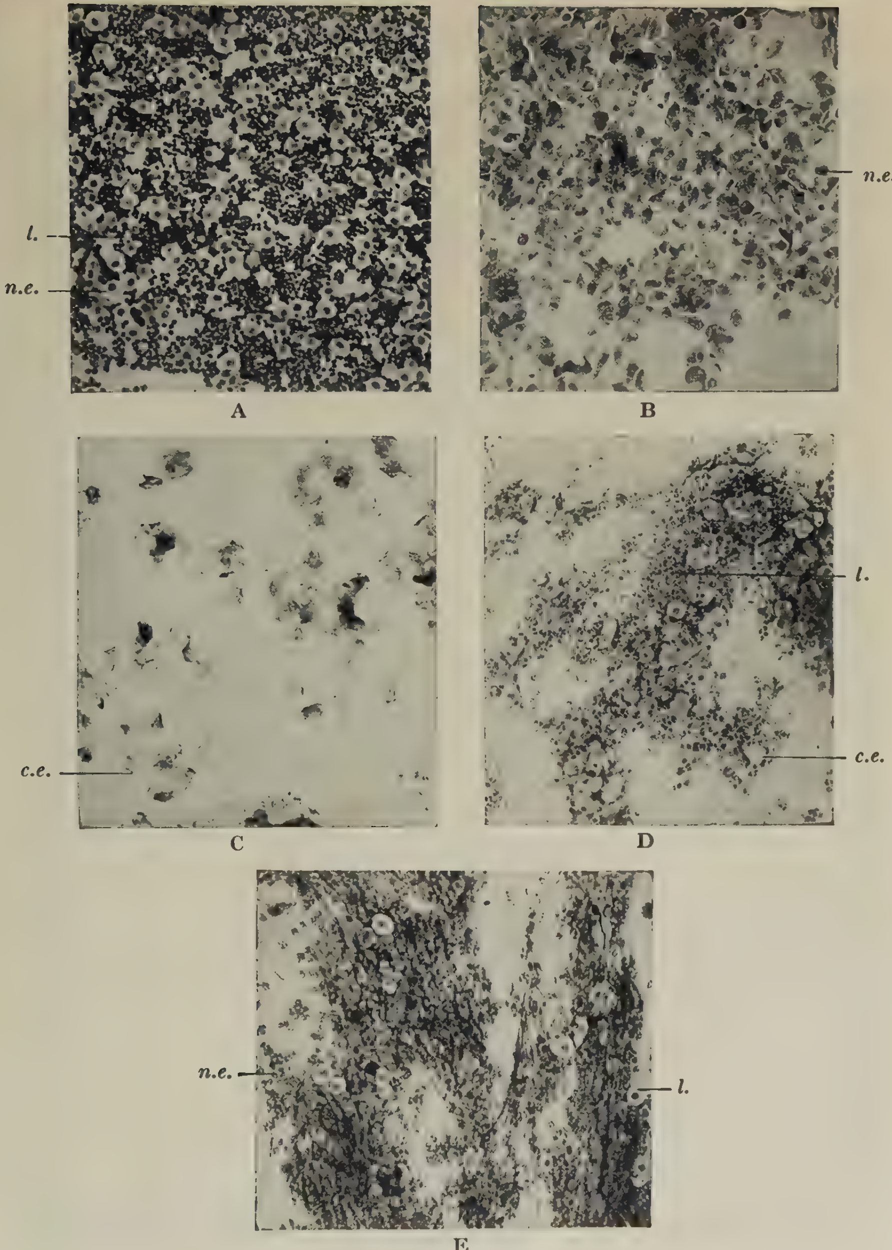


Fig. 10. 8. **Vaginal Smears** of the Mouse, at various stages of the Œstrous Cycle.

A. *Di-œstrus* : the smear consists largely of leucocytes (*l.*) and nucleated epithelial cells (*n.e.*).

B. *Early Œstrus :* the leucocytes have disappeared, leaving only the nucleated epithelial cells.

C. *Œstrus :* the epithelial cells have become cornified and non-nucleated (*c.e.*).

D. *Early Di-œstrus :* leucocytes have reappeared, and there are fewer cornified epithelial cells.

E. *Pregnancy :* the smear consists of leucocytes and epithelial cells, mixed with much mucus, and, during the later stages, blood.

In *pseudo-pregnancy* the smear is similar to that in di-œstrus, with the addition of much mucus. (From Parkes' " Internal Secretions of Ovary.")

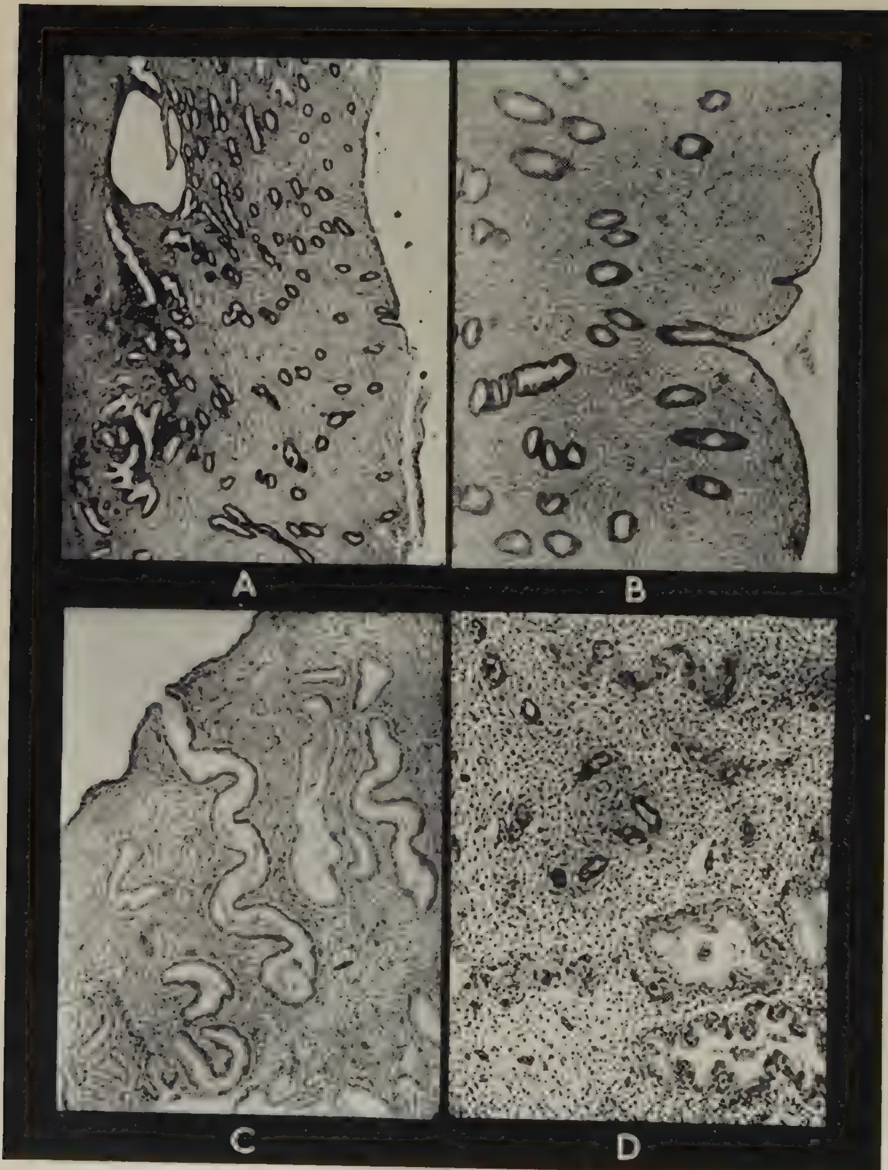


FIG. 10. 9. Human Endometrium.

A. Early proliferative (cycle day 7) ; small tubular glands (× 15).
B. Late proliferative (cycle day 13) ; the glands are enlarging (× 23).
C. Early secretory (cycle day 17) ; the glands have become cork-screwed and contain secretory material within the lumina ; sub-nuclear vacuolation is present (× 40).

D. Late secretory (cycle day 27) ; the glands (bottom right) have a florid epithelium, now showing signs of exhaustion ; stromal œdema is present ; spiral arterioles (top left) are prominent (× 67). (P. E. Hughesdon.)

hormones produce in the uterus. These have been studied experimentally in spayed monkeys, some of which have a menstrual cycle similar to that of women.

The basal layer of the endometrium lying directly on the myometrium is about 1 mm. thick and contains arterioles and the ends of the uterine glands embedded in fibrous tissue. In the spayed animal this is covered by a superficial layer, about 0·5 mm. thick, consisting of connective tissue, capillaries and the necks of the glands ; it is surmounted by low columnar or cubical epithelium. Administration of œstrogen causes hypertrophy of the superficial layer to a depth of 3 or 4 mm., the blood vessels becoming dilated and increased in length, often in the form of spirals, the glands also increasing in length and the epithelium becoming high and columnar (Fig 10. 9). The water content of the uterine tissues is increased. This is precisely the state of the normal endometrium in mid-cycle just before ovulation.

Sudden withdrawal of œstrogen results in dehydration and shrinking of the endometrium with compression and partial occlusion of the spiral blood vessels, leading to very sluggish blood flow and some stasis in the dilated capillaries. The supply of oxygen to the mass of tissue probably becomes insufficient, for the endometrium, instead of involuting gradually as it does when œstrogen is slowly withdrawn, shows punctate hæmorrhage and necrosis and is cast off with some blood into the uterus. The débris here undergoes autolysis and is discharged into the vagina. This process is very similar to that of normal menstruation.

The interval between œstrogen deprivation and hæmorrhage varies considerably but averages about a week. The slower the deprivation (up to the point where it is too slow to be effective in causing hæmorrhage), the longer is the interval. A period of maximal secretion lasting from a week to a fortnight in the middle of the cycle, with a fairly rapid rate of rise and fall, would give a typical spacing of menstrual flows and allow for the mucosa to have reached its highest by about the tenth or eleventh day. Although no one has yet made a reliable direct determination of the rate at which the ovaries secrete œstrogen, fluctuations in the urinary œstrogen output accord with the postulate, except that there is a second increase in œstrogen excretion in the second part of the cycle, coincident with the phase of activity of the corpus luteum.

In anovulatory menstrual cycles, in which follicle ripening culminates in atresia instead of ovulation, the mechanism underlying menstruation probably corresponds closely to that just described. But such cycles are relatively uncommon : ovulation is the rule and is followed by the formation of a corpus luteum, whose activity must therefore be accounted for. When stimulated by luteotrophic hormone of the anterior pituitary (p. 295, below) it secretes a substance, *progesterone* (Fig. 10. 6), which has been isolated from it and can replace it after surgical removal. Progesterone can be shown in many ways to reproduce effects normally associated with the presence of the corpus luteum.

Injection of progesterone into the œstrogen-primed castrate monkey

causes the epithelium of the uterus to increase in height and become active. The first recognisable change is the development of a secretory vacuole found beneath the nucleus of each gland cell (*subnuclear vacuolation*). The secretory material is then discharged into the lumen of the gland, which becomes distended, and because the glands have proliferated to a greater extent than the surrounding stroma, they become tortuous and saw-edged in section. Sub-epithelial connective tissue cells become rounded and more closely packed, making a compact superficial zone in the endometrium resembling decidua. These cells respond by proliferation to irritation of the endometrium (*e.g.* by the presence of a hair) and make a tumour appropriately called a *deciduoma.* Beneath the compact zone is one of œdema and vasodilatation. The basal fibrous layer remains unchanged. This is the state of the normal endometrium just before menstruation in the human cycle. Sudden withdrawal of progesterone injections will precipitate a breakdown of the endometrium similar to that already described, except for the more advanced condition of the mucosa. This may occur in spite of the œstrogen injections being maintained constant.

Samples of endometrium have been transplanted to the anterior chamber of the eye in monkeys and studied by direct observation. The sum of evidence suggests that œstrogen and progesterone are the two props of the endometrium, progesterone having little or no influence on it in the absence of preliminary treatment by œstrogen. If either of the two is suddenly withdrawn, the endometrium will break down with hæmorrhage. If they are gradually withdrawn, it may subside without necrosis. Since it has been shown that each may have an independent effect, it must be presumed that menstruation in normal, ovulatory cycles is due to the coincident effects of the withdrawal of both hormones. Possibly the basic cycle is determined by fluctuations in œstrogen concentration (and reactivity of the endometrium) while the exact timing depends on degeneration of the corpus luteum.

There is no doubt that the chief function of the corpus luteum is to prepare the endometrium for gestation. During the luteal part of the cycle the endometrium is called *progestational*, and the physiological counterpart of the deciduoma is the formation of decidua in response to implantation of an ovum in it. The decidual cells and glands contain glycogen. There are lipoid granules in the epithelium and stroma and the whole probably nourishes the embryo before its circulatory system is established.

In the human, *pregnanediol* (Fig. 10. 6) is one of the most easily measured metabolites of progesterone. Its excretion in the urine is significantly increased during the luteal phase of the cycle and, to an increasingly greater extent, during pregnancy.

The luteal phase in animal cycles is variable. In the rat and mouse it is negligible, the corpora lutea being practically nonfunctional in the absence of copulation. If this occurs, a long luteal phase follows which, failing pregnancy, is called *pseudopregnancy.* In dogs and guinea-pigs, the phase is well marked even in the absence of copulation but is also often

called pseudopregnancy. In rabbits, cats and ferrets, ovulation is not spontaneous and a luteal phase is entirely absent unless copulation takes place. Copulation with a vasectomised buck in the rabbit, however, yields such a well-developed progestational endometrium that its aspect at a certain time in pseudopregnancy has been used as a standard for the assay of progestational substances.

Some animals do and some do not bleed during the breakdown of progestational endometrium and some may bleed at another point in the cycle, *e.g.*, œstrus. Uterine bleeding in animals is in itself of little help in correlating their cycles with those of women.

The time of ovulation in the cycle has been investigated in man and the primates in various ways.

(1) By the study of changes in the endometrium. Using a small curette, it is comparatively easy to remove fragments of the endometrium which can then be examined histologically. The proliferative phase in the endometrium is associated with maturation of the follicle and, following ovulation, the secretion of progesterone by the corpus luteum causes subnuclear vacuolation as the earliest change indicating ovulation. The transition from proliferative to secretory phase takes place from about the fourteenth to sixteenth days of a twenty-eight-day cycle.

(2) By examining the ovaries at operation. Over a period of years, many corpora lutea resulting from recent ovulation have been seen at operation and their apparent age has been correlated with the time in the cycle at which they were seen.

(3) By recovery of ova from the uterus and Fallopian tubes. During gynæcological operations the tubes and uterus have been washed out and newly shed ova recovered. Allowing for time taken to travel along the tubes, the time of ovulation has then been estimated.

(4) By recording electrical changes. It is said that, at the time of ovulation, there is a marked potential difference between an electrode placed on the symphysis pubis and one in the vagina.

(5) By the accumulation of records of cycles in which coitus took place once only on a known date. The frequency with which fertilisation and pregnancy followed can then be related to the day of the menstrual cycle on which coitus occurred. Although fertilisations have been recorded on almost every day of the cycle, the greatest frequency is found on the twelfth, thirteenth and fourteenth days of the cycle.

(6) By the study of changes in the vaginal smear. Although in women these are not so distinct as those found in rodents, characteristic changes can be seen in smears taken daily. The proportion of epithelial cells showing pyknotic nuclei increases rapidly up to the time of ovulation, at which time leucocytes are typically absent. Immediately afterwards they suddenly reappear, the proportion of pyknotic cells decreases, while clumping and folding of the cells—characteristic effects of progesterone—are seen.

(7) By the study of changes in cervical mucus. This increases in quantity just prior to the time of ovulation, becomes clear and watery and free from cells and is said then to contain glucose. Shortly after

ovulation the mucus becomes tacky and cellular and, incidentally, more or less non-receptive to spermatozoa. If the mucus is allowed to dry on a glass slide, typical fern-like crystals of sodium chloride can be seen under the microscope in mucus just prior to and at the time of ovulation. Ferning disappears usually shortly after ovulation.

(8) By observing changes in the *basal temperature*, that is, the body temperature taken shortly after waking in the morning. The same changes are seen whether the temperature is taken in the mouth or the rectum. In ovular cycles there is characteristically a rise of approximately 1° F. shortly after the time of ovulation with a temperature plateau persisting until the onset of the next period. This temperature shift is caused by progesterone and the time of transition from low to high level probably corresponds fairly closely with ovulation. It is often said that a fall in temperature occurs on the day of ovulation ; this finding is inconstant and without significance.

(9) By studies of urinary œstrogen excretion. These reveal that there are characteristically two peaks of œstrogen excretion during the cycle : a sharp one occurring about the twelfth to fourteenth day and believed to coincide with the day of ovulation, and a more sustained peak during the luteal phase of the cycle.

From results obtained in these ways, it has become clear that ovulation most commonly occurs on days 12, 13 or 14 of cycles which average twenty-eight days. When the cycles are longer, it is the pre-ovular phase which is lengthened, the post-ovular phase remaining relatively constant at fourteen or fifteen days (though in some women with long cycles the post-ovular phase is also somewhat prolonged). It is often stated that ovulation may occur on any day in the cycle. Without qualification, this statement is only a half truth. A woman who has regular cycles ovulates on much the same day in each cycle, but since there are few women who do not occasionally have either a longer or shorter cycle than usual, there are few women who *never* ovulate other than during the expected " fertile phase." Though isolated coitus on what should have been an infertile day does occasionally lead to conception, it must be understood that, had conception not occurred, the period would have been abnormally early or late as the case may be.

Gonadotrophins

The anterior pituitary (adenohypophysis) is believed to elaborate three *gonadotrophic* principles. One of these, the *follicle stimulating hormone* (F.S.H.), causes ripening of follicles but does not by itself lead to ovulation. Another is the *luteinising hormone* (L.H.) which causes luteinisation of the follicles, large and small alike. The thecal and granulosa cells increase in size so that the follicle becomes solid and the ovum disappears. Vascularisation occurs and the cells come to contain lipoid material which is often yellow in colour. The resulting body is a corpus luteum. In a hypophysectomised animal, a carefully balanced dosage of F.S.H. and L.H. will cause ovulation, but neither alone will do so : it results, probably, from the supplementary action of L.H.

when F.S.H. has ripened the follicle. The combined activities of F.S.H. and L.H. also are needed to cause secretion of œstrogens from the ovarian follicles. In the male, the same gonadotrophic hormones are found, and it is often stated that F.S.H. causes spermatogenesis and L.H. the secretion of androgens by the interstitial cells of the testis. (Hence L.H. is sometimes called the *interstitial cell stimulating hormone*—I.C.S.H.) However, since it has proved extremely difficult to obtain pure preparations of these two gonadotrophins, it cannot be assumed with certainty that in their actions on the testis either is wholly effective in the absence of the other. A third gonadotrophic hormone secreted by the anterior pituitary gland is called *luteotrophic hormone* (Lt.H). The presence of this appears to be necessary for secretory activity by the corpus luteum, leading to the production of progesterone. Lt.H. is considered to be identical with *prolactin*, a pituitary hormone involved in lactation.

The gonadotrophic hormones of the pituitary, therefore, have a double function—the maintenance both of the production of sperm and ova (gametogenesis) and of the various internal secretions of the gonads as is indicated diagrammatically in Fig. 10. 15 (p. 309, below). Another gonadotrophic hormone, called *chorionic gonadotrophin*, is found in the serum and urine of pregnant women ; it is produced by the trophoblast of the human embryo. This hormone serves to maintain the corpus luteum of pregnancy, though whether it does this by a direct action on the corpus luteum or by stimulating the pituitary gland to secrete L.H. and Lt.H., is still uncertain.

If this hormone is injected into hypophysectomised rats or mice, it luteinises their ovaries and will also cause, in the males, enlargement of the seminal vesicles and ventral lobe of the prostate gland (due to secretion of androgen by the interstitial cells of the testes—which suggests that it may well have a direct action on the gonads). Given to intact rodents, it causes ovulation and this provides the basis for the Aschheim-Zondek and Friedman tests for pregnancy. Other properties which are utilised for pregnancy diagnosis tests are the ability to cause ovarian hyperæmia in certain strains of rats, external ovulation in *Xenopus laevis*, the South African clawed toad, and the expulsion of spermatozoa into the cloaca in certain amphibians. Human chorionic gonadotrophin is fairly readily obtained and so is used in clinical medicine (though its usefulness is distinctly limited). The only other readily obtained gonadotrophin for clinical use is that which is found in the serum (not the urine) of pregnant mares and is known as serum gonadotrophin, or PMS (pregnant mare's serum). This substance has a predominantly follicle-stimulating action. Neither of these substances acts in the human as pituitary gonadotrophins evidently do and they are not able, with any certainty, to bring about ovulation. This is in contradistinction to their effects in certain animal species, in which they can be used regularly and with certainty to produce ovulation, or super-ovulation (the ripening and discharge of an abnormally large number of eggs).

The Climacteric

The reproductive period in women is commonly much shorter than the total life span. The main reason for this is the continual loss or degeneration of primordial follicles in the ovary and the absence of any

provision for their replacement. When no more ova remain, reproductive ability obviously ceases. The endocrine activity of the ovary, however, does not end in this abrupt manner and so there is a phase of waning function. This phase begins before the periods actually cease and continues for some time afterwards. The whole phase is properly called the *climacteric*; but the term is sometimes confined to the interval between the beginning of ovarian failure and the *menopause*, or cessation of the periods.

The age at which the climacteric sets in varies greatly in different women in much the same way as does that at which the *menarche* (the time of the first period) occurs; in about 80 per cent. of women the menopause sets in between the ages of 42 and 55. Various factors—genetic constitution, general health, social influences and so on—no doubt help to determine it and diseases of various sorts directly affecting ovarian physiology can accelerate it. It has often been supposed that there is a relationship between the age of the menarche and that of the menopause but accurate statistics suggest that this is not true, the menopause occurring at much the same time in women who began their periods early as in those who started late.

When the number of ovarian follicles has become significantly reduced, although more or less regular cyclic activity may continue for some time, ovulation does not occur in every cycle. Later ovulation ceases altogether and by this time some irregularity of the cycles has usually become apparent. Anovular cycles may continue for some time, the bleeding becoming more infrequent and scantier, and eventually ceasing altogether. In some women, however, the periods may become heavy and prolonged before they cease at the menopause. In yet others, cessation of menstruation is an abrupt change without any antecedent disturbance of menstrual rhythm.

The gradual lessening of endocrine activity during the climacteric leads to many secondary changes. Foremost among these are regression of the genital organs, uterus, vagina, vulva and breasts. But in many women, the changes are almost imperceptibly slow and long after the menopause little definite evidence of general regression may be seen. Further secondary effects of ovulation failure are changes in the activity of the other endocrine glands. These are most important for a full understanding of the physiology of the climacteric. Because of the progressive failure of ovarian response to pituitary gonadotrophic stimulation, the modifying action of the ovarian hormones on the activity of the pituitary becomes less and less. With the reduced inhibitory effect of œstrogen, the production of pituitary gonadotrophin (mostly of follicle-stimulating type) increases, so that in post-menopausal women a marked excess can be detected in the urine. There is probably also an increase in the secretion of thyrotrophic and adrenocorticotrophic hormones by the pituitary gland. These find responsive target organs and so hyperactivity of the thyroid and adrenal cortex may ensue. It is probably the combination of falling œstrogen level and increased output of thyroid and adrenal hormones which is responsible for most

of the untoward effects which may be experienced by women at this stage. The increased activity of the adrenal cortex, however, tends to restore the endocrine balance by taking over some of the functions of the ovary. Thus, there is little doubt that most of the œstrogen which circulates in post-menopausal women, when the ovaries have become quite functionless, is secreted by the adrenal cortex. The œstrogen and the other adrenocortical hormones produced in slightly increased amounts probably serve to depress the excessive pituitary activity and so reduce the thyroid overactivity. The tendency to develop a considerable growth of hair on the face seen in some women after the menopause is directly due to the increased output of androgens by the hyperactive adrenal and the decreased level of circulating œstrogen.

Coitus, or Copulation

This is the act of union whereby the male deposits spermatozoa in the genital tract of the female. Coitus is attended by excitement which culminates in the *orgasm*—a paroxysm of sensation largely contributed by sensory elements in the glans penis and accompanied by the ejaculation of semen. The degree of sexual excitement experienced by females varies considerably in different sub-human species but it is doubtful if, in most, orgasm is experienced at all. Among women, great variability of orgastic experience is encountered ; in some it is intense, in others totally absent.

The nervous basis for coitus is a spinal reflex and the act can occur after section of the spinal cord in the dorsal region, when there is complete absence of sensation. The two essential parts of the act are erection, which enables the penis to be inserted into the vagina, and ejaculation. *Erection* is the result of distension with blood of the venous sinuses of the corpus spongeosum and of the corpora cavernosa whose resistant fibrous capsules then render the penis hard and rigid. This is brought about by dilatation of the helicine arteries of the penis, as a result of which inflow of blood into the corpora cavernosa increases, while through compression of veins the outflow of blood is hindered. Associated with this is relaxation of the smooth muscles in the trabeculæ of the fibrous tissue. Stimulation of the pelvic nerves (second, third and fourth sacral segments) initiates erection and their section abolishes it. Stimulation of sympathetic fibres from the lumbar region is said to constrict the vessels of the penis and make it flaccid. The afferent side of the reflex arc conveys sensory impulses from the penis but superimposed upon the basic reflex arc are the effects of impulses from the higher nervous centres, by means of which many other stimuli, such as sight, smell, sound, as well as the results of purely cortical activity, such as thought, memory and so on, can cause erection. At the same time, many stimuli acting through the association areas of the brain can exert an inhibitory effect on the erection reflex, either preventing its occurrence or abolishing it once it has begun. This inhibitory mechanism is held responsible for most cases of impotence.

It also provides the means whereby some control of a voluntary nature can be exercised so that, for example, whereas in some circumstances various stimuli may evoke erection, in others where erection would be undesirable it can be deliberately prevented. Involuntary inhibition of an erection in progress may result from disturbing influences of all sorts occurring at an inopportune moment.

A further important factor affecting erection is the presence of male sex hormone (testosterone, p. 284, above). Erections occur in boys long before puberty, in eunuchs (men whose testes have been removed) and in eunuchoids (men whose testes have never developed properly) ; hence, testosterone is not essential for erection. On the other hand, many eunuchoids complain of infrequent and imperfect erections which are rendered normal by appropriate treatment with male hormone. Moreover, if given in excessive doses, testosterone may induce a state of more or less continuous erection (called priapism). On the other hand, the administration of testosterone to men who are impotent but who secrete normal amounts of male hormone is almost invariably without any effect whatsoever. It would seem, therefore, that testosterone facilitates the normal erection reflex, reducing the threshold of the stimuli necessary to excite it ; it is powerless, however, to overcome the effects of inhibition exerted by the higher centres and it is quite clear that these inhibitory stimuli are prepotent since, coming in circumstances which have already excited erection, they can abolish it.

When the stimuli which excite erection are sufficiently intense and sufficiently prolonged, they set in train a remarkable series of nervous and muscular effects, culminating in orgasm and ejaculation. Pulse and respiratory rates increase and blood pressure rises ; there is a general development of muscular tension throughout the body and rhythmic movements, especially of the pelvic region, occur and increase in speed. At the climax, or orgasm, tensions are released, ejaculation occurs and the body then rapidly returns to its normal state. Ejaculation, like erection, is brought about primarily by a spinal reflex and once the reflex has been set in motion it is beyond the reach of voluntary inhibitory control. In this it is unlike the subjective accompaniments, since these can be enhanced by influences acting on the higher centres of the brain. Preceding the actual ejaculation, the stimuli, which will eventually produce it, cause reflex increased secretion of the accessory sex glands so that, in some men, clear fluid (mainly derived from the glands of Littré) may escape from the urethral meatus. The discharge of impulses from the spinal centre eventually causes rhythmic contractions of the vasa deferentia, seminal vesicles and prostate, thereby expelling the contained spermatozoa and accessory secretions. The seminal fluid so formed is ejected from the urethral opening in a series of spurts, varying in number from two or three to perhaps a dozen. The first spurt is usually devoid of spermatozoa, being composed chiefly of secretions from the urethral glands and Cowper's glands. An intermediate fraction of the ejaculate is rich in spermatozoa, the remainder consists

mostly of secretions from the seminal vesicles and prostate gland and may be almost devoid of spermatozoa. The smooth muscles involved in ejaculation are stimulated by excitation and paralysed by section of their sympathetic supply. Section of the presacral nerve in man therefore causes sterility.

The forceful nature of the first part of the ejaculation is due to contractions of accessory muscles which compress the root of the penis. At the same time the sphincter muscle at the base of the bladder contracts thus preventing the back-flow of seminal fluid into the bladder. Very occasionally, apparent failure to ejaculate (in spite of a normal orgasm) may be caused by the failure of this muscle to contract appropriately ; the condition can be diagnosed by examining the urine voided shortly after an orgasm when live spermatozoa will be found.

In the female, erectile tissue in the vulva and around the vagina also becomes engorged during sexual stimulation and the secretions of the greater and lesser vestibular glands are discharged, providing lubrication during coitus.

Fertilisation

Fertilization normally occurs in the ampulla of the Fallopian tube. At this stage, the ovum is still surrounded by numerous follicle cells derived from the cumulus oöphorus, embedded in intercellular substance composed largely of *hyaluronic acid.* Spermatozoa carry an enzyme called *hyaluronidase* by means of which they are enabled to penetrate through the intercellular substance, so as to reach the zona pellucida. It was at one time thought that fertilisation could not take place until a sufficient number of spermatozoa had accumulated in the vicinity of the ovum, still enclosed within the cumulus, so that the hyaluronidase concentration of the surrounding fluid would permit dissolution of the cumulus ; then spermatozoa could reach the ovum itself. This view is now known to be untrue, since, by means of phase contrast microscopy, the presence of a fertilising spermatozoon has been demonstrated within the vitellus still enclosed inside an intact cumulus.

When the fertilising spermatozoon enters the vitellus, the ovum undergoes a second division and the second polar body is extruded. At this stage, the nucleus of the ovum is called the *female pronucleus* and that of the spermatozoon, the *male pronucleus.* The two pronuclei proceed to unite, followed immediately by the first mitotic division of the embryonic nucleus, to give the first two daughter cells.

As soon as one spermatozoon penetrates the zona pellucida, a chemical transformation is propagated over the surface of the zona, preventing the entry of other spermatozoa. This is called the " block to polyspermy " and normally prevents the entry of more than one spermatozoon. In certain circumstances, however, this block may fail and multiple penetration of the ovum by spermatozoa may result. But when this happens, the normal sequence of events leading to fusion of the male and female pronuclei does not occur, so that fertilisation fails.

Sex Determination and the Differentiation of the Sexes

The factors determining the sex of an individual are primarily genetic. The nuclei of the ovum and spermatozoon each contain only half the number of chromosomes (the *haploid* number) present in the remaining cells of the body but by their union they form a cell whose nucleus has the normal (*diploid*) sum. Chromosomes in all cells except the gametes are therefore paired and so are the genes which they carry. For every maternal gene which affects, for instance, eye colour, there is in the complementary chromosome a corresponding paternal gene which influences it in the same or a different way. In the latter event, the final eye colour will be decided by the " dominant " gene of the pair and the " recessive " gene will be powerless until the next generation. Apart from this kind of genic inequality, the half set of chromosomes in an ovum is similar to that in other ova and to that in half the spermatozoa of members of the same species. In the other half of the spermatozoa, one chromosome is modified. This is called the Y chromosome and is smaller than the alternative chromosome, which is called X, X and Y being the " sex chromosomes " (Fig. 10. 10). Half the unions between ova and spermatozoa result, therefore (theoretically), in cells having nuclei with two exactly paired sets of chromosomes.

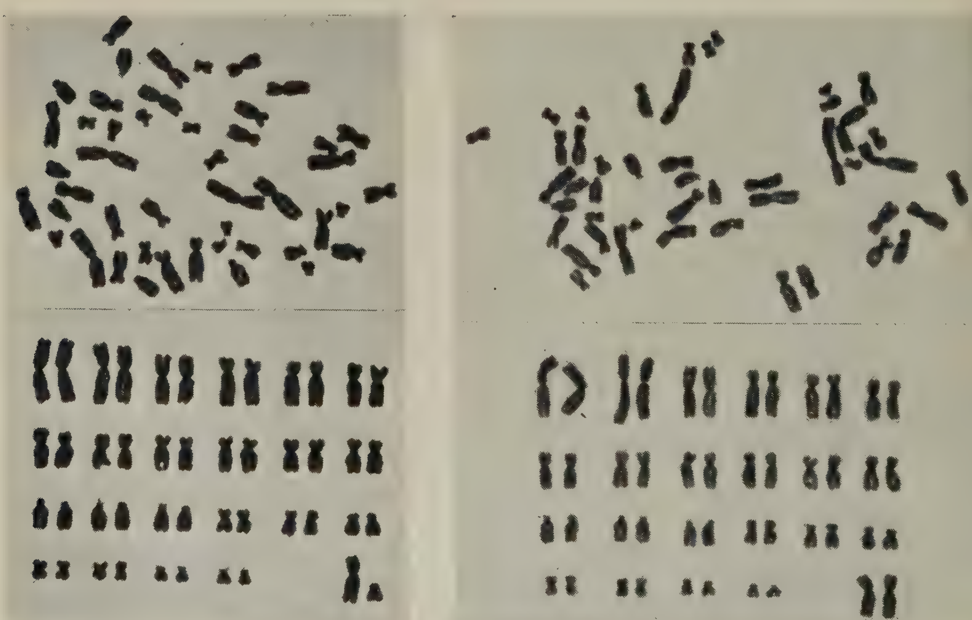


FIG. 10. 10. Chromosomes of Normal Human Cells.

Left : male. Right : female.

The 44 autosomes of each cell have been selected and arranged in pairs in descending order of size, leaving the two sex chromosomes, XY in the male cell and XX in the female cell, as shown at the bottom right-hand corners. (From photographs kindly provided by Prof. L. S. Penrose.)

These cells divide and differentiate to form individuals bearing ova, *i.e.*, females. The other half of the union yields cells in whose nuclei one member of one pair of chromosomes differs slightly from its fellow. These cells divide and differentiate to form individuals producing spermatozoa, *i.e.*, males. Females are said to be *monogametic* and males *digametic* (in moths and birds the same principle holds good but the female is the digametic member).

Genetic Sex Differentiation. A survey of the animal kingdom shows that *secondary sexual characters* (variations in parts of the body other than in the gonads and accessory organs which characterise animals as male or female) are governed by two agencies, respectively genetic and chemical. The first is found in its purest form in the insect world.

We have seen (Fig. 10. 10) that the combination of two similar sex chromosomes (XX) yields a female and of dissimilar sex chromosomes (XY) a male. The accepted explanation is that the X chromosome carries a factor making for femaleness which, when doubled, is sufficiently strong to balance the tendency of the combined remaining chromosomes (*autosomes*) to produce maleness. When single, as in the XY combination, it is not strong enough. In insects, the form of all sex variable parts of the body emerges according to the state of this balance. For example, the shape of a wing is not determined solely by the coincidence of a *single* gene in a paternal chromosome with the corresponding gene in the corresponding maternal chromosome, but by the combined effect of several genes scattered through each complementary half set of chromosomes. The gene, or genes, which can influence wing shape and which reside in the second sex chromosome have the " casting vote " on sex form.

Sometimes in a genetically female zygote (XX) a fault occurs in the first mitosis of the fertilised egg and the X chromosomes of one of the daughter cells are altogether lost. Since each daughter cell gives rise to one-half of the body, all the cells in, say, the left half are, sexually XX and in the right half, OO. The result of this accident is an insect in which the left half of the body in all its sex variable parts is female and the right half male. Other degrees of genic imbalance which give a range of intersexual forms can be produced experimentally in insects.

Chemical Sex Differentiation. The findings of embryologists, many in lower vertebrates and many depending on experiments difficult to perform and to interpret, suggest the following generalisation.

In vertebrate embryos, the primitive gonad is bi-potential, the cortical part being capable of developing into an ovary and the medullary part into a testis. The influence of the genetic factor is exerted at a very early stage, causing the suppression of the medullary part with development of the cortex in the case of females and of the reverse situation in the case of males. It will be appreciated that a failure of the normal balance between " femaleness " sponsored by the X chromosomes, and " maleness " by the autosomes, can, in certain circumstances, lead to a disturbance of the above-described process so that either both cortex and medulla develop to form an ovo-testis (as is found in some

true hermaphrodites) or a testis may develop on one side and an ovary on the other, or various other possible combinations (as have also been described in human true hermaphrodites) may occur. Moreover, it is possible for an ovary to develop in a genetic male or a testis in a genetic female in the same kind of way and such gonadal anomalies are believed to exist in certain humans. Such gonads are defective to a varying extent.

Further experimental studies, in which the gonads in the " indifferent " stage have been removed from embryos or have been destroyed *in situ* by X-irradiation, have demonstrated that, whereas the removal of the gonads in a genetic female does not disturb the normal development of female assessory sex organs, if those of a genetic male are removed, the individual develops as though it were a female. From this it has been concluded that the embryonic testis secretes a substance which is necessary for the development of the male accessory organs (*i.e.*, Wolffian duct derivatives, penis and scrotum) and for the suppression of the female counterparts, while in its absence the female structures (*i.e.*, Müllerian duct derivatives and vulva) develop and the male structures are suppressed. There is experimental evidence that treating embryos at a sufficiently early stage with sex hormones may modify the development of the primitive gonad. One of nature's best-known experiments of this kind is found in cattle when the circulations of twins of genetically opposite sex communicate with each other during the early indifferent stage of development. The male twin develops normally (possibly because its ovarian rudiment disappears before it can be stimulated) but the " female " twin, a so-called " free-martin," is extensively modified. Sterile testes and male genital ducts are formed, the ovaries and female ducts being suppressed. The external genitalia are indeterminate, usually of a rudimentary female type, though the clitoris may be enlarged.

Although differentiation of mammalian sex characters is thus vested mainly in the sex hormones, the genetic foundation on which these work is often subject to modification. This is well seen in the plumage of some birds but is also apparent in man. Many traits, such as cephalic index, presence or absence of palmaris muscle, and height, about which we say " most men are taller than most women," and so forth, are thought to be fundamentally genetic in origin.

Intersexuality occurs in man in three forms : true hermaphroditism, chromosomal intersexuality and male and female pseudo-hermaphroditism. The first is rare and implies the presence of gonadal tissue of both sexes in the same individual. It has been mentioned above. Chromosomal intersexes are individuals with sex-chromosome aberrations. One group, with XO constitution (45 chromosomes only), consists of hypogonadal " females " with dysgenetic gonads (" chromatin negative Turner's syndrome "). Another, with XXY constitution (47 chromosomes) consists of hypogonadal males with abnormal testes (" chromatin positive Klinefelter's syndrome "). Reduplication of sex chromosomes may also give rise to XXX,XXXYY and similar anomalies, most of the individuals so affected being mentally defective as well as showing sexual abnormalities. A further group consists of individuals believed to be " mosaics," some nuclei having one sex chromosome

constitution and others another. In the pseudo-hermaphrodites, the sex of the gonad is indicated by the adjective " male " or " female," while the external genitalia and the secondary sex characters partake to a greater or less extent of the nature of those associated with the opposite sex.

In *male pseudo-hermaphrodites*, testes are present but are usually undescended, and although they show apparently normal interstitial cells, spermatogenesis does not occur. The external genitalia are mal-developed to a greater or less extent. Thus, the urethral meatus opens under the glans

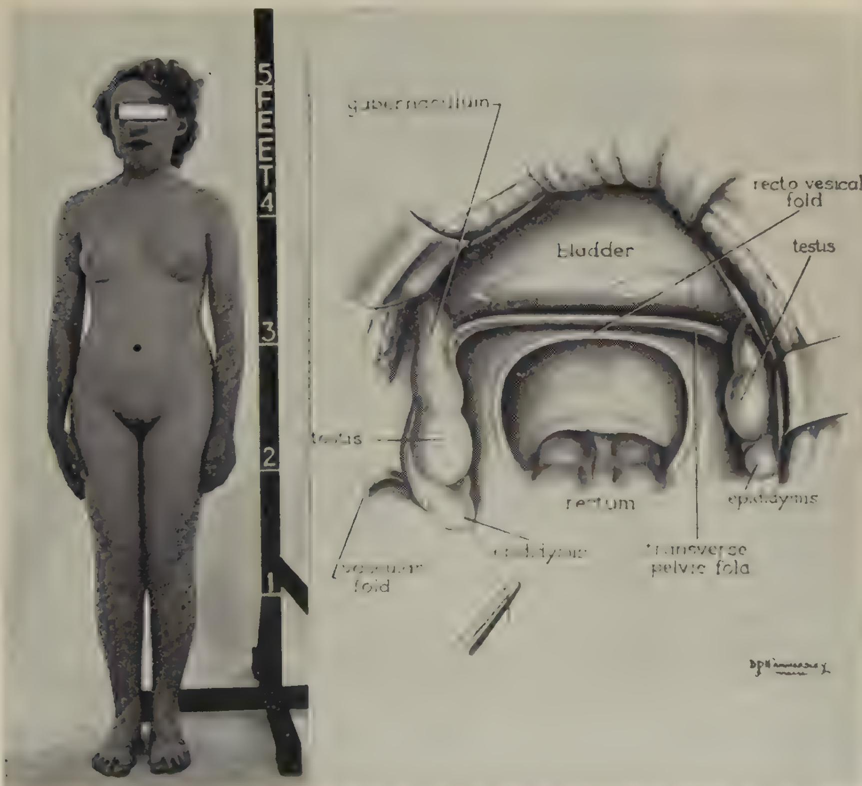


FIG. 10. 11. Male Pseudo-hermaphroditism.

A patient, 17 years old, with " testicular feminisation." Well developed breasts and external feminine characteristics (in fact, won a beauty competition) but no pubic or axillary hair and infantile vulva. At laparotomy, no internal female genitalia, but undescended testes found. Sex chromatin negative. (Case of Dr. C. N. Armstrong.)

penis or at the base of the penis or several centimeters behind in the perineum (hypospadias). Sometimes the penis is so poorly developed as to resemble merely an enlarged clitoris. The internal genital organs appropriate to the female are sometimes found in a rudimentary form. In some individuals, flat male breasts are found but in others, well-developed breasts of female type are found. Indeed, in one group of these patients (sometimes referred to as the " testicular feminisation syndrome "), the individuals appear to be normally developed females except that pubic and axillary hair are usually absent and the vagina is short and ends blindly, no uterus being present

(Fig. 10. 11). In some intersexual individuals it may be impossible to determine, on clinical grounds, whether they are male or female pseudo-hermaphrodites except on the basis of hormone studies, biopsy of the gonads or nuclear sex studies.

The commonest cause of *female pseudo-hermaphroditism* (Fig. 10. 12) is a genetically determined disturbance of adrenal function arising in embryonic

Fig. 10. 12. Female Pseudo-hermaphroditism.

A patient aged 24 who complained of genital abnormalities and primary amenorrhœa. Bodily configuration male with fairly well developed muscles, no breasts and a narrow pelvis. Phallus measured 3 cm. and urethra opened at its base. No vaginal orifice. Very high excretion of adrenal steroid metabolites, reduced to normal by prednisone (which suppresses pituitary ACTH secretion). Laparotomy revealed ovaries, uterus and tubes and a cervix opening into a vagina, access to which was gained by incising the perineum, so constructing a vaginal orifice. The phallus was amputated. With continued prednisone treatment, menstruation began within a month, has continued normally and the patient has conceived and delivered a normal infant. (Case of Dr. G. I. M. Swyer.)

life and leading to the production of excessive amounts of androgenic steroids by the adrenal cortex. These, though not interfering with the development of the ovary or of the internal female genital apparatus, cause great enlargement of the clitoris, so that it resembles a hypospadiac penis. They may also produce varying degrees of vaginal mal-development, and cause the early appearance of hirsutism and of masculinisation. In a small proportion of these individuals, the adrenal defect leads to excessive sodium

loss, so that, unless treated, the infant dies within the first few weeks of life. Very rarely, a non-adrenal form of female pseudo-hermaphroditism is encountered, in which the only defect appears to be enlargement of the clitoris. This has been observed to result from treatment of the mother with male hormone during pregnancy but it is also known to occur occasionally after treatment with progesterone-like substances or even when there has been no treatment at all. In these cases, it is believed that the embryo must have been abnormally sensitive to the slight androgenic action of the maternal or administered progesterone-like bodies.

Pregnancy

As previously mentioned, fertilisation of the ovum takes place in the ampulla of the Fallopian tube. During the next three or four days, the fertilised ovum is transported through the Fallopian tube by the co-ordinated muscular contraction of its walls and thereafter, for a further five or six days, it lies free in the uterine cavity. It is at this period that secretory activity of the endometrial glands is at its highest and presumably the material secreted serves for the nourishment of the developing blastocyst. By the end of this time, however, the trophoblast has appeared on the surface of the blastocyst and by means of the proteolytic enzymes secreted by this tissue, the embryo is enabled to burrow into the endometrium in which it becomes implanted. The trophoblast secretes *chorionic gonadotrophin* which now enters the maternal circulation and prevents degeneration of the corpus luteum. Because of the continued secretory activity of this organ, the endometrium develops into decidua and menstruation does not occur. Thus begins the *physiological amenorrhœa* of pregnancy. Not infrequently, however, a little bleeding occurs at the time of implantation ; it is usually less than that of a normal period but may sometimes lead to confusion about whether pregnancy has or has not taken place. (There are, of course, many other causes of suppression of menstruation, particularly notable among which are emotional factors which are believed to act through the hypothalamus and anterior pituitary gland.) Chorionic gonadotrophin passes readily through the maternal kidneys and appears in the urine. It has sometimes been detected as early as the twenty-sixth day of the cycle—that is, before the day on which the period would have occurred had not conception taken place. However, the amounts at this stage are very small and it is not until about a fortnight after the first " missed period " that pregnancy tests depending upon its detection in the urine become reliable. By this time, physical examination may already show fairly unmistakable signs of pregnancy, though in many women the changes may not be definite until two or three weeks later than this. The duration of pregnancy is conventionally measured from the first day of the last menstrual period and is approximately 280 days.

From about the sixth week onwards, for a variable period usually lasting four to six weeks, the mother often experiences nausea and vomiting, particularly in the morning. This *morning sickness* may vary from a trifling inconvenience to a most distressing experience (when it

is called *hyperemesis gravidarum*). Its cause is not known with certainty but a possible explanation is that it is in the nature of a hypersensitivity reaction to chorionic gonadotrophin which is, after all, a foreign protein, being partly of paternal origin. During this time, other symptoms which may be experienced are pain or enlargement of the breasts and irritability of the bladder, causing increased frequency of micturition.

The uterus hypertrophies to accommodate the growing embryo, the fundus rising above the level of the symphysis pubis at about the twelfth week and reaching the xiphisternum at about the thirty-sixth week, when it drops somewhat owing to the head of the fœtus sinking in the pelvis. At about the sixteenth week, movements of the fœtus can be appreciated by the mother (quickening) and thereafter become more and more obvious. Symptoms and signs of increased intra-abdominal pressure may include digestive disturbances, some œdema of the legs, dilatation of the superficial veins of the legs, dyspnœa and palpitation. Other obvious changes are in the size and vascularity of the mammary glands which hypertrophy and from which a clear fluid can be expressed in the latter half of pregnancy. Pigment is deposited in the areola of the nipple, which turns from pink to brown in the first pregnancy, and also in the linea nigra of the abdomen.

Nitrogen is retained by the mother in excess of the amount needed by the fœtus and this is true also of calcium if sufficient is supplied in the diet. The adequacy in the diet of substances like calcium, iron and vitamins needs more careful supervision than its calorific value. There is retention of water by the mother, frequently a decrease in the erythrocyte count of the blood, and an increase in leucocytosis towards term. Other blood changes are lipæmia, cholesterolæmia and a diminished alkali reserve. The cardiac output progressively increases in the later months, probably to keep pace with the quantity of blood which has to be passed through the placenta to maintain the oxygen supply of the fœtus.

Although the fœtal and maternal bloods do not mix, they are in intimate relation with each other (Fig. 10. 13). Oxygen, taken in by the mother's lungs, diffuses from the maternal blood across the thin capillary endometrium and attenuated syncytium of the placental villi into the fœtal blood. This transfer of oxygen is greatly facilitated by the fact that the affinity of fœtal blood for oxygen is greater than that of maternal blood (Chapter 3, p. 79). The fœtal blood will take up oxygen from the maternal blood, and its percentage saturation will increase, even though the partial pressure of oxygen in the fœtal blood must be less than that in the maternal blood if oxygen is to diffuse across. Glucose and aminoacids likewise diffuse freely across from mother to fœtus, and fats probably do so much more slowly. Carbon dioxide and most of the other metabolic products of the fœtus diffuse in the other direction, to be removed by the mother's lungs, or excreted by her kidneys. Most proteins and other substances of large molecular weight and most formed elements are stopped at the placental barrier. Thus, for example, thyroxine and other non-protein hormones can

penetrate but not the thyrotrophic hormone of the anterior pituitary. Some antibodies, notably some of those concerned in the agglutination of the red blood cells, can penetrate—a matter of some importance, as discussed in Chapter 21—and also some viruses (for example, smallpox) and possibly some bacteria. There is little doubt that in the

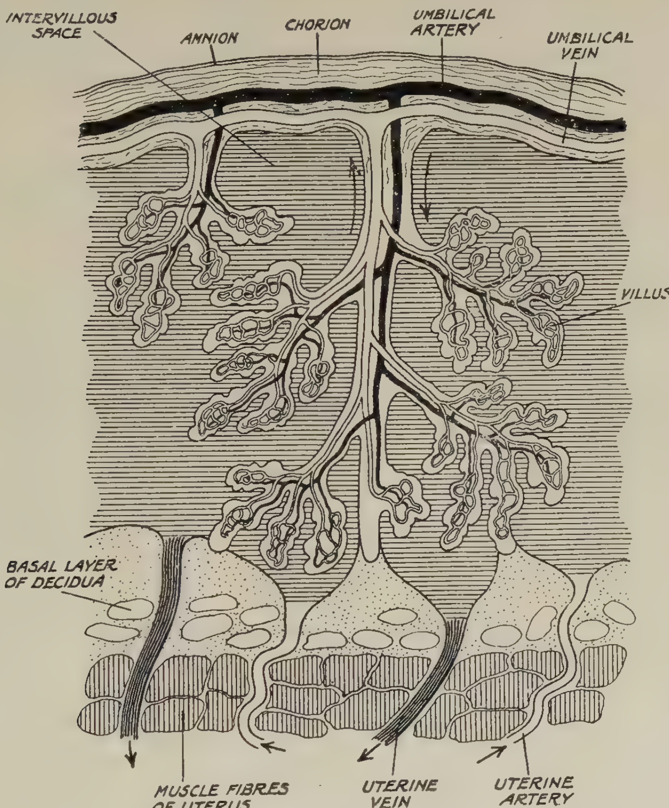


FIG. 10. 13. Diagram of the relations between the Fœtal and Maternal Bloods in the Placenta.

De-oxygenated blood flows from the fœtus in the umbilical artery, and passes into the villi, which are imbedded in the placental decidua of the uterus ; in these it takes up oxygen and foodstuffs from the maternal blood, and loses carbon dioxide and waste products ; it then returns to the fœtus in the umbilical vein. The intervillous space is well supplied with blood by the uterine arteries, and forms part of the maternal circulation. (After Gray.)

first three months or so of pregnancy, when the syncytium and Langhans layer are fully formed, the trophoblast contains enzymes which enable it to attack the maternal tissues and stores, for example, of glycogen, in the neighbourhood and plays an active part in nourishing the embryo. Fructose, for example, is formed from glucose and passes into the fœtal blood but not into that of the mother.

The placenta is now known to be important in other ways, though it is not clear how much this is due to the fœtal and how much to the maternal part. In mice, rats and monkeys, where the placenta is fully established, the fœtus may be removed without disturbing the placenta and the latter will remain and be delivered at the time for normal parturition. Meanwhile, the extra-uterine signs of pregnancy persist, though if the placenta is also removed prematurely they disappear. This suggests that pregnancy is by no means a hand-to-mouth adjustment to the demands of the fœtus but a maternal syndrome with a definite duration presided over by the placenta.

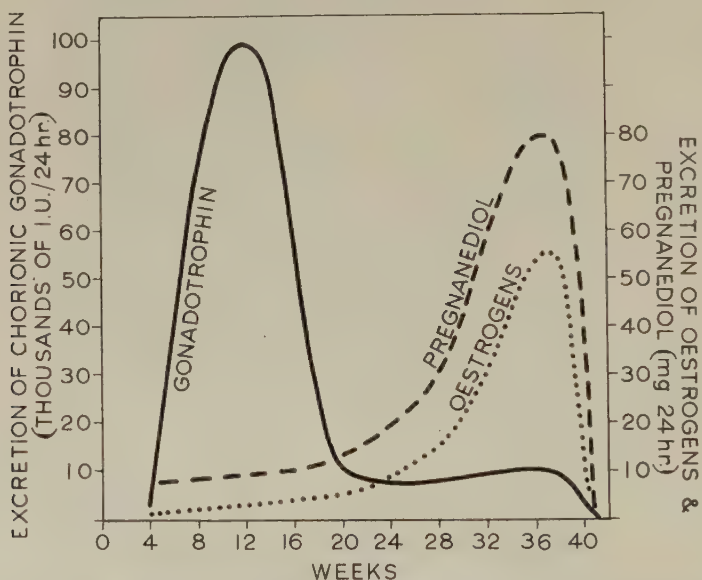


FIG. 10. 14. Urinary excretion of chorionic gonadotrophin, pregnanediol and œstrogens during pregnancy.

The urinary excretion of chorionic gonadotrophin, œstrogen and pregnanediol in pregnancy follows the curves shown in Fig. 10. 14. The daily output of œstrogen may reach more than a thousand times, and that of pregnanediol fifteen times, that of the maximum found during the menstrual cycle. Chorionic gonadotrophin is found in the urine only during pregnancy (or when certain tumours containing chorionic tissue are present). Its urinary output increases rapidly to a peak at about the seventieth day of pregnancy and then equally rapidly falls to a level which is maintained throughout the rest of pregnancy, a tendency to rise being shown during the last trimester. It is in accordance with experimental observation that increasing amounts of chorionic gonadotrophin are needed to maintain a functional corpus luteum and by the time chorionic gonadotrophin is falling, the placenta has begun to take over the functions of the corpus luteum as a source of progesterone and

œstrogens. In many animals, removal of the corpora lutea terminates pregnancy, but in man pregnancy continues provided the corpus luteum is not removed before about the twelfth week or, if earlier, adequate amounts of progesterone are given. There is little doubt that the increasing amounts of œstrogen and progesterone secreted during pregnancy are responsible for the changes in maternal anatomy and physiology which have already been mentioned and it is probable that a nice adjustment between the various groups of hormones is important for the

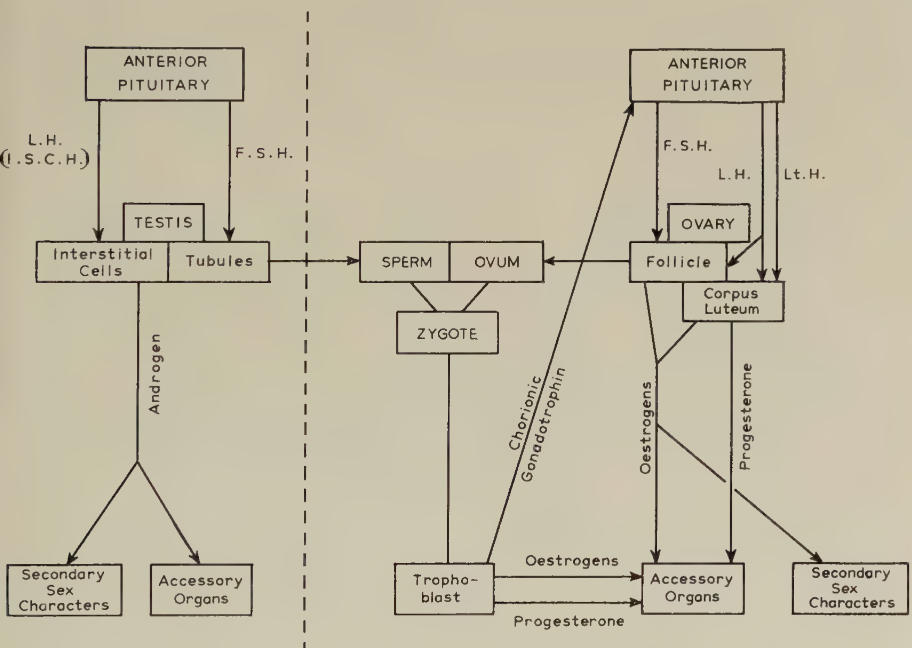


FIG. 10. 15. Diagrammatic summary of the inter-relationships of the principal hormones concerned with reproduction.

L.H. = luteinising hormone ; F.S.H. = follicle stimulating hormone ; I.C.S.H. = interstitial cell stimulating hormone ; Lt.H. = luteotrophic hormone.

maintenance of a normal pregnancy. Although progesterone is usually regarded as essential for pregnancy, it is difficult to say if it is in fact any more important than are the œstrogens, though one reason generally advanced for its indispensability is its depressing effect on uterine motility and tone, enabling the organ to " give " before the growing fœtus.

Relaxation or increase in length of the ligaments of the sacro-iliac joints makes these more mobile than usual and a similar mobility of the symphysis pubis may occur. The changes are probably important during parturition, rendering the pelvic canal less rigid, but are not so

marked as those which occur in some animals (for example, guinea-pig and mouse). A second luteal hormone, "relaxin," is concerned in these changes in some animals.

A general outline of the somewhat complicated inter-relationships of these hormones—pituitary, ovarian and placental—is given in diagrammatic form in Fig. 10. 15. The important part that is played by the pituitary gland, both adenohypophysis and neurohypophysis, in the reproductive processes, including, as we shall see below, parturition and lactation, will be discussed further in Chapter 11.

The Fœtal Circulation. The arrangement of the blood vessels in the fœtus differs from that in the same animal after birth, owing to the fact that the lungs and alimentary canal are functionless and the whole of the nutrition has to be obtained from the mother through the placenta. The lungs are shunted out of the circulation in two ways: first, oxygenated blood passes into the child from the placenta in the umbilical vein. This mixes with de-oxygenated blood from the legs and arrives at the heart in the inferior vena cava, whence it is deflected by the Eustachian valve through the foramen ovale directly into the left side of the heart. From the aorta this relatively well oxygenated blood passes to the brain and the systemic circulation generally. Secondly, blood from the brain enters the right side of the heart in the superior vena cava and passes from the pulmonary artery into the aorta through the ductus arteriosus, entering at a point beyond that at which the vessels supplying the brain leave. The brain therefore is not supplied with blood that has in large part just left it. It should be noted that none of the blood in the fœtal circulation is truly arterial except that in the umbilical vein; the blood supplied to the head and upper limbs is more arterial than that supplied to the abdominal organs (most of which, incidentally, are almost functionless) and the lower limbs, since it contains a large proportion of blood derived directly from the placenta while the rest of the body receives blood that has mainly been through the vessels of the upper regions already.

Parturition. In the later months of pregnancy, abdominal palpation reveals a distinct hardening of the uterus at intervals, undoubtedly due to contraction of the muscle. These contractions are often felt by the mother but do not become unpleasant in character until the onset of labour, or parturition, when the pains which may then accompany them are often the first symptom. Simultaneously there is commonly a small discharge of mucus tinged with blood from the vagina. It is difficult to say whether the contractions of labour are painful on account of anoxia of the muscle (a well-known cause of muscular pain) or, if so, in what way they differ from the earlier painless contractions of pregnancy. It may be assumed that they are more powerful. Labour is traditionally divided into three parts. In the *first stage*, the fœtal head fitting the brim of the pelvis compresses the membranes and traps some of the amniotic fluid below it, the bulging "bag of waters" being palpable by a finger being passed through the dilating cervix. This fluid medium transmits the pressure exerted by the uterine contractions

evenly to the cervix which eventually dilates sufficiently to allow the head of the fœtus to pass. At some time during this stage the membranes rupture and amniotic fluid is discharged.

Full dilatation of the cervix marks the beginning of the *second stage* which ends with the complete delivery of the fœtus. The severing of the umbilical cord should be delayed until all the placental blood has drained into the baby, to whom it presents six months' supply of iron. During the whole of the first and second stages the pains marking rhythmic contraction of the uterus become progressively more frequent or severe and more prolonged. They may be separated by fifteen to thirty minute intervals in the first stage and by only a few seconds in the second which is, however, relatively short in a normal delivery (Fig. 10. 16). The first stage varies greatly in duration, diminishing with the number of previous pregnancies from, say, sixteen hours to six.

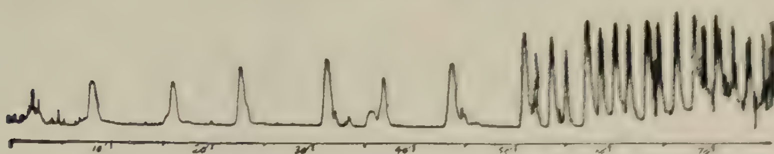


FIG. 10. 16. Record of the **Intra-uterine Pressure** of a Woman in **Labour.**

The contractions are relatively infrequent up to the fifty minutes' mark, when the first stage ends and the second begins ; they immediately become more frequent and more violent. The head of the child was born after seventy-five minutes, at the moment indicated by the sudden fall in the pressure below the previous minima, and the body, just after the last rise in pressure. (Bourne and Burn.)

The *third stage* occupies about half an hour, during which the placenta is detached and expelled. The raw bleeding surface thus left is rendered harmless by a firm contraction of the uterus which is said to be enhanced by putting the baby to the breast. If it fails to occur, hæmorrhage may be severe. Methods of stimulating contraction artificially are to administer the *oxytocic principle* of the posterior pituitary gland subcutaneously or ergometrine intramuscularly after the placenta is expelled ; more recently, however, it has often become the practice to give ergometrine intravenously or subcutaneously, together with hyaluronidase (a " spreading factor "), at the time of the birth of the fœtal head to guard against post-partum hæmorrhage. The only drawback to this procedure is that there is an increased frequency of retained placenta which then may have to be delivered manually—an expedient straightforward enough in hospital but fraught with danger if done outside it.

Many investigators maintain that the posterior pituitary secretion plays a physiological *rôle* in normal labour, and stimulation of the hypothalamic region produces uterine contractions in rabbits when all connection between the two parts, save the blood stream, has been severed. It would appear that the oxytocic principle, which undoubtedly has a powerful stimulating effect on the uterus, is more than a mere active tissue extract. Parturition may occur in many animals from which

the pituitary gland has been removed ; in cats and guinea-pigs, however, normal parturition does not occur if the nerve supply to the posterior pituitary from the hypothalamus has been interrupted. Nevertheless, in the human, normal parturition has been known to occur in patients with diabetes insipidus (Chapter 11).

It has been suggested that the uterus is sensitised to the action of oxytocin by the presence of œstrogen (which does not have this effect in all animal species) and that oxytocic activity is suppressed by the presence of progesterone. If, moreover, the production of progesterone by the placenta were to diminish rapidly at the end of pregnancy, the uterus might well become more active without the aid of the pituitary. The evidence, however, is still too uncertain to justify further discussion.

Distension of the uterus is another obvious stimulus which, it has been considered, on reaching a certain threshold, may cause the onset of labour, but this is ruled out as an essential factor by the delivery at term in some animals of placentæ of negligible bulk, the fœtuses having previously been removed. Several other stimuli for the onset of labour might be postulated and the cumulative effect of numerous stimuli is a possibility. Parturition can take place after section of the spinal cord in the dorsal region and in dogs even after destruction of the lumbo-sacral cord. There is some evidence, however, that in the human being, lumbo-sacral centres are necessary.

Lactation

The mammary glands progressively enlarge during pregnancy and animal experiments show that this is almost certainly due to the combined effects of œstrogen and progesterone. Although in some animals, for example the guinea-pig, œstrogen is capable of bringing about complete development of the mammary gland, in most it serves only to enlarge the duct system while the alveoli proliferate under the influence of progesterone (Fig. 10. 17). However, œstrogen and progesterone appear to be ineffective in producing mammary development in hypophysectomised animals and it is now generally agreed that growth hormone and the adrenocortical hormones are essential for the process. (Hypophysectomy, of course, leads to adrenal atrophy through loss of adrenocorticotrophic hormone.) For the secretion of milk, it is generally agreed that prolactin (which is probably the same as luteotrophic hormone) is necessary.

The natural occurrence of lactation does not follow the same pattern in all animals. In the mouse, for instance, secretion into the alveoli seems to run parallel with their development, so that at full term the gland is in a fully secreting condition and producing milk. This happens, however, under the influence of the placenta and not of the pituitary as can be shown by removing the latter during pregnancy. Immediately parturition has taken place in these hypophysectomised mice, the mammary glands involute and no milk is obtained by the litter. Continuation of secretion is therefore dependent on the anterior pituitary. This is also true in the human being, as in shown by the fact that

necrosis of the pituitary gland, which may follow severe post-partum hæmorrhage (Sheehan's syndrome), is associated with absence of lactation. On the other hand, in the rhesus monkey, from which the fœtus has been taken and in which the placenta has been left *in utero*, mammary development is less than normal, even though the pituitary

FIG. 10. 17. Experimental development of mammary gland. (After Turner, 1932.)

(1) Mammary gland of male mouse (× 8).
(2) Mammary gland of male mouse after four rat units of œstrogen per day for thirty days. This corresponds to the development in a virgin adult female (× 4½).
(3) Mammary gland of spayed rabbit after twenty days' injection of progesterone. This had been preceded by œstrogen treatment, but no corpus luteum effect is visible (× 1).
(4) Mammary gland of spayed rabbit after treatment with œstrogen and progesterone simultaneously. This corresponds to the stage naturally reached at mid-pregnancy (× 1).

has not been molested. No hard and fast rule can therefore be applied to the development of secretory activity in all animals.

In women, a clear fluid called *colostrum* accumulates in the mammary glands in the latter part of pregnancy and true milk does not appear until some three days after parturition. Secretion is then profuse and dramatic and is in the nature of a definite event rather than the climax of a slowly developing process. The rapid clearance of œstrogen from the maternal circulation (it falls to a very low level after parturition)

is held responsible for the sudden release of prolactin from the pituitary which is turn causes the onset of profuse milk secretion. Relief of tension by suckling becomes imperative but if this is not allowed to take place, lactation will cease in a few days. The act of suckling (or otherwise emptying the breast) initiates nervous reflexes which lead (*a*) to a further secretion of milk and (*b*) to the expulsion from the depths of the mammary gland of the milk already secreted. The formation of the milk by the secretory cells is induced by the reflex release of prolactin from the anterior pituitary ; the expulsion of milk is due to the reflex release of the oxytocic hormone of the posterior pituitary acting on the myo-epithelial cells within the gland. Both parts of the pituitary are thus concerned with lactation at this stage. It is usual to wean children by the time they are six to nine months old and although many cases of lactation for a number of years are known, it is unlikely that suckling would prolong it indefinitely.

Clinically, it has been found that some failures in lactation are due to failure of milk ejection and that an injection of oxytocin at the commencement of suckling may considerably increase the milk yield to the baby. It is probable that these cases are of psychological origin, for one treatment with oxytocin may prove sufficient. When a woman does not wish to feed her baby at the breast and lactation must be suppressed, it is usual to give œstrogens which are commonly but erroneously supposed to inhibit lactation. They serve to prevent the engorgement of the breasts which occurs when the latter are not emptied, but if suckling or emptying of the breasts continues along with the œstrogens, lactation persists. It is, in fact, the failure to empty the breasts which eventually brings about suppression of lactation. Testosterone preparations are also effective in preventing engorgement.

The Properties of Milk. Milk is a white, opaque fluid, usually very slightly acid in reaction, containing as its chief constituents sugar (lactose) and a protein (caseinogen) in solution and fats in suspension. It also contains two other proteins—lactalbumin and lactoglobulin—in smaller amounts, and salts. It is interesting to note that the ratio of calcium to sodium and potassium in the ash is considerably greater in milk than it is in blood, although it approximates much more closely to the ratio in the whole animal ; a similar correspondence between the composition of milk and that of the whole animal is also found in the ratio of phosphate to chloride.

If allowed to stand, milk is liable to undergo fermentation by the *bacillus lacticus*, with the formation of lactic acid by hydrolysis of the lactose. The consequent increase in acidity results in the coagulation of the caseinogen on boiling (curdling) or, eventually, in its coagulation at room temperature. On standing, moreover, the greater part of the fat globules rises to the surface with the formation of cream. This process can be hastened by the use of a centrifuge, the pattern used in commercial dairying being known as a cream separator. Violent agitation of cream, as occurs in a churn, results in a phase reversal ; cream is a suspension of oil and fat globules in a watery medium, while butter

is a suspension of watery droplets in an oily medium. The solidity of butter results from its di-phasic nature, just as does that of the grease used for lubricating bearings.

Chemically, the constitution of milk is admirably adapted for the needs of the growing animal and it contains all those substances enumerated in Chapter 7 as necessary for the maintenance of health. The diet of a normal infant has a greater energy value per unit body weight and contains considerably more fat than that of a fully grown person. The chemical composition of the milk is not the same in all animals, however, and the Table below shows that cows' milk, for example, contains considerably less lactose and more caseinogen and salts (except those of iron) than does human milk. This accounts in large measure for the difficulty that is sometimes found in raising babies on cows' milk, or preparations made from it, when, for any reason, the mother is unable, or unwilling, to nurse it.

TABLE 10. 1

The Percentage Composition of Milk

	Water	Proteins		Fat	Lactose	Salts
		Caseinogen	Albumin and Globulin			
Human milk .	88·5	1·2	0·5	3·3	6·0	0·2
Cow's milk .	87·1	3·0	0·5	3·7	4·8	0·7

CHAPTER 11

DUCTLESS GLANDS

Ductless glands secrete substances into the blood stream. These substances travel to and control the activities of other tissues which are usually remote from the secreting gland and sometimes widely distributed in the body. Such chemical messengers are called *internal secretions* to distinguish them from secretions which, like bile or the digestive juice from the pancreas, are transmitted through ducts. More commonly they are called **hormones** (from a Greek word meaning " I arouse to activity "), the name suggested by W. M. Bayliss and E. H. Starling after their studies in 1902 on how secretin excites pancreatic secretion. The ductless glands discussed in this chapter and in that on Reproduction are often called the endocrine organs and their study is the rapidly developing branch of physiology known as endocrinology.

Hormones may be extracted from most of the ductless glands by crushing or boiling them in water. The study of the actions of tissue extracts, however, is in the first place a branch of pharmacology, for it must not be supposed that the action observed, however specific or potent, necessarily represents the normal function of the gland. One might as well imagine that the effect of digitalis on the vertebrate heart was a normal function of the glucoside in the foxglove from which it had been extracted. The assertion that a particular action is a normal function of a ductless gland depends on the correspondence between the consequences of excision of the organ, and replacement by injection of extracts, or of implantation of gland tissue.

To avoid confusion, it is essential to distinguish between the specific and the non-specific actions of tissue extracts. All crude tissue extracts contain histamine and produce capillary dilatation and a fall of blood pressure when injected intravenously (Fig. 11. 1) ; in exceptional circumstances, this depressor effect may be masked by the action of a sufficiently potent pressor substance. A specific depressor effect can, therefore, only be distinguished from the non-specific action of histamine, if the effective substance in a tissue extract can be isolated and purified. The actions of acetylcholine and of very small doses of adrenaline, shown in Fig. 11. 1, are instances of this kind. A pressor action, on the contrary, is readily recognised as a specific effect of the extract which induces it, as exemplified by extracts of the posterior lobe of the pituitary gland (Fig. 11. 13), and of the adrenal medulla in all but minute doses (Fig. 2. 12, p. 55).

Criteria of an Endocrine Gland. Before an organ can be classified as an endocrine gland it must be shown to fulfil most of the following criteria.

(1) *Histological structure.* On microscopic examination the ductless

glands reveal the presence of glandular cells, no ducts, but a copious blood supply. The cells frequently possess granules (anterior pituitary cells) or lipoid filled vacuoles (adrenal cortex, corpora lutea, interstitial cells of the ovaries or testes). The size and cytological appearance of the cells may be related to the state of activity of the gland.

(2) *Hypofunction.* Removal of the gland in an experimental animal

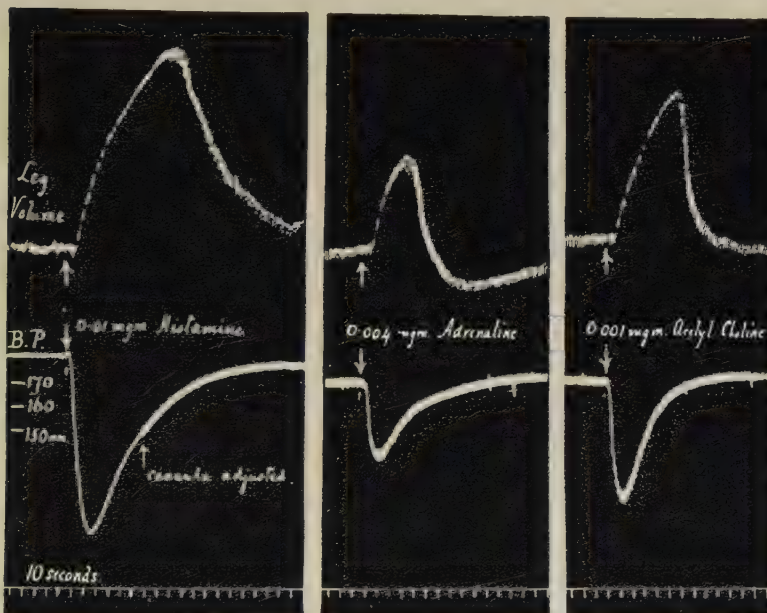


FIG. 11. 1. The Action of **Histamine, Adrenaline** (in minute concentration) and **Acetylcholine** on the Arterial Pressure and the Volume of the Denervated Limb of a Cat.

A larger dose of adrenaline (say, 0·05 mg.) would have given the more characteristic vasoconstrictor effect, *i.e.* rise in arterial pressure and diminution in limb volume. (Dale and Richards.)

or man, or disease of the gland in man, produces a state of " hormonal deficiency."

(3) *Replacement therapy.* The state of hormonal deficiency may be corrected by administration of extracts of the gland.

(4) *Transplantation.* Endocrine glands usually function normally if transplanted to a distant site in the body. Successful transplantation demonstrates that any factors maintaining and regulating the activity of the gland do so *viâ* the blood stream, and that the gland exerts its action by secretion into the blood stream. The first demonstration of endocrine activity was made by Berthold in 1849 when he showed that a transplanted testis in the cock could maintain the comb of the bird. Unfortunately from the point of view of human therapeutics, tissue

transplanted from one species to another, or even from one individual to another, may not " take " and grow (see Chapter 21).

(5) *Hyperfunction.* Administration of large amounts of extracts of a gland or some types of tumour of the gland result in a condition of " hormonal excess."

(6) *Assay of venous blood from the gland.* Hormone activity has not yet been detected in the venous blood from all the ductless glands (*e.g.* anterior pituitary gland) ; when it can be detected, however, it shows clearly that endocrine activity is occurring in the particular structure, especially if it can be shown that the amount of hormone in the blood varies in a regular manner in varied physiological states.

General Control of the Endocrine System. The endocrine organs co-operate with the nervous system in co-ordinating the activities of the various parts of the body ; and accordingly they are influenced by changes in the external and internal environments and react appropriately. For example, the secretion of insulin by the pancreas depends on the concentration of sugar in the blood ; seasonal changes of light, temperature and food supply may profoundly modify the activity of the ovary or testis (Chapter 10) ; changes in environmental temperature affect the activity of the thyroid gland (Chapter 20) ; stressful or noxious procedures excite the activity of the adrenal cortex and medulla, but inhibit the thyroid and gonads ; and changes in the osmotic pressure of the blood, tactile stimuli to the mammary glands and possibly coitus activate the posterior pituitary gland.

The endocrine system is unsuited to effect changes with the speed of nervous reflexes. Some ductless glands, particularly those which are controlled by a secretomotor nerve supply (*e.g.* adrenal medulla, posterior pituitary gland), secrete hormones which exert their actions within seconds after liberation into the blood stream, and are then quickly destroyed in the blood or tissues, or rapidly excreted. Other glands are themselves controlled by hormones (*e.g.* thyroid, gonads). These react more slowly to any appropriate stimulus and secrete hormones which are more stable in the blood stream and exert their effects slowly and over a more prolonged period. The secretomotor nerve supply to the posterior pituitary gland, and that to the adrenal medulla, are derived ultimately from the hypothalamus. This part of the brain also controls the anterior pituitary gland, and since the trophic hormones from this gland maintain and regulate the activity of the thyroid, adrenal cortex, ovary and testis, it is clear that the basal region of the brain determines, either directly or indirectly, the activity of the major part of the endocrine system. The hypothalamus may be affected both by afferent nervous stimuli and by changes in the composition of the blood. The effects of these factors are probably integrated in the hypothalamus, which thereby co-ordinates and regulates the activity of both the autonomic nervous system (Chapter 15) and the endocrine glands. It may be said that this part of the central nervous system is intimately concerned with the maintenance of a constant internal environment.

Many hormonal actions have been considered in previous chapters under their appropriate systems : the secretions of the gastro-intestinal mucosa (gastrin and secretin, for example) in Chapter 5 ; of the pancreas (insulin) in Chapter 6 ; of the gonads (androgens and œstrogens, for example) and of the pituitary body (gonadotrophins) in Chapter 10. This chapter will deal with some of these in rather more detail, and with some hormonal actions which have not yet been considered. Various organs have been numbered among the ductless glands in the hope that a suitable function would be found for them ; some, like the carotid body, are now known to have quite a different kind of function (Chapter 2), while others, like the pineal body, can as yet have no function confidently assigned to them.

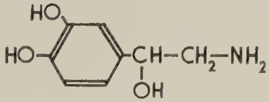
THE ADRENAL GLANDS

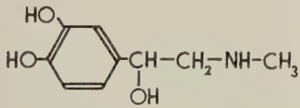
The two adrenal glands are situated immediately above the upper poles of the kidneys. Each consists of two histologically distinguishable portions : (1) The cortex, which is yellow in colour and stains darkly with silver nitrate or fat soluble dyes such as Sudan III ; (2) the medulla, which is smaller and darker, stains brown with chromates, and is consequently referred to as " chromaffine tissue." The cortex arises from the cœlomic epithelium, while the medulla arises from the ecto-derm, in close association with the sympathetic nervous system. The two portions of the gland are thus distinct in origin and appearance, and, as their functions do not seem to be related, it will be convenient to consider them separately.

The Function of the Adrenal Medulla

Complete removal of the medulla in the presence of functioning adrenal cortical tissue can be achieved by enucleation of both adrenal glands from their capsules. Functional cortical tissue regenerates from the remains of the capsule, but medullary tissue does not regenerate. Once cortical tissue is re-established the animal suffers from no untoward symptoms and exhibits no lowering of the arterial pressure.

Extracts of the medulla, however, are of great physiological potency. Their actions have been traced to substances chemically related to tyrosine, and named **adrenaline** and **noradrenaline.** They have the following formulæ :—





The proportion of noradrenaline that can be extracted from the adrenal medulla varies with the species. Rabbits' adrenals contain little if any

noradrenaline, human and cattle adrenals about 20 per cent., and those of the cat about 50 per cent. of the combined total content of adrenaline and noradrenaline. The naturally occurring form of adrenaline is the lævo-rotatory isomer, and this is about twelve times as active as the dextro-rotatory form. Similarly L-noradrenaline has many times (twenty-five to thirty times) the activity of D-noradrenaline. Rapid destruction of adrenaline by oxidation, and slightly less rapid destruction of noradrenaline, occurs in solution at pH 7 when exposed to air. Owing to its instability, administration of adrenaline by mouth has little effect. Intravenous injection of adrenaline has a short-lived effect, since most of it is destroyed by the tissues in a few minutes. A more prolonged action follows subcutaneous injection because absorption is delayed by the vasoconstriction produced by the adrenaline itself. Small amounts of adrenaline and noradrenaline are normally excreted in the urine.

It was originally suggested that noradrenaline was only a metabolic precursor of adrenaline. However, it is now established that both substances may be present in adrenal venous blood, and it has been claimed that electrical stimulation of different regions of the hypothalamus, which is well known to activate the adrenal medulla, may cause liberation of a secretion which is either preponderantly adrenaline or noradrenaline. This makes it likely that in normal circumstances these substances may be liberated independently. Indeed, it has been suggested that emotional excitement, well known to cause adrenal medullary activation, liberates largely noradrenaline in *aggressive* animal species or individuals, and adrenaline in defensive, *passive* forms.

In general it may be said that adrenaline and noradrenaline act on the organs in the body innervated by the sympathetic system and have the same action on them as has stimulation of the sympathetic nerves (with the exception of the sweat glands). This parallelism is closer for noradrenaline than adrenaline ; noradrenaline is probably the chief natural sympathetic transmitter at most post-ganglionic nerve endings. The functions of the sympathetic nervous system will be discussed further in Chapter 15 on the autonomic nervous system.

The sympathomimetic actions of adrenaline and noradrenaline on the cardio-vascular system were discussed in Chapter 2. Both produce a rise in arterial pressure, which is greater after section of the vagi. Adrenaline has an augmentor and accelerator action on the heart, a constrictor action on the arterioles of the skin and kidneys and a dilator action on those of the skeletal muscles (in man) and the coronary system of the heart. The vasodilator effect is often more easily evoked than the vasoconstrictor effect by a very minute concentration of adrenaline, so that in suitable animals an overall vasodilatation may be produced by injection of minimal doses of adrenaline with a consequent fall in blood pressure (see Fig. 11. 1). Noradrenaline has little or no vasodilator action and has little direct action on the heart, but is more potent than adrenaline in producing vasoconstriction. The musculature of the alimentary canal and of its outgrowths, such as the bronchi, is relaxed by both adrenaline and noradrenaline, though the

sphincters may contract; the pupil is dilated. Administration of adrenaline produces a discharge of glycogen from the liver, with consequent hyperglycæmia and possibly glycosuria (Chapter 6), and an increase in the rate of discharge of the adrenocorticotrophic hormone from the anterior pituitary gland. Noradrenaline is less potent than adrenaline in both these respects.

The action of adrenaline persists after section of the sympathetic nerves, even if time has been allowed for complete degeneration of their peripheral ramifications; the heart and the pupil are, in fact, more sensitive to adrenaline when this has been done; therefore, it does not act on the nerve endings (Chapter 15, p. 463). The action of adrenaline may be prevented by drugs, such as ergotoxine and apocodeine, which also interrupt the influence of the

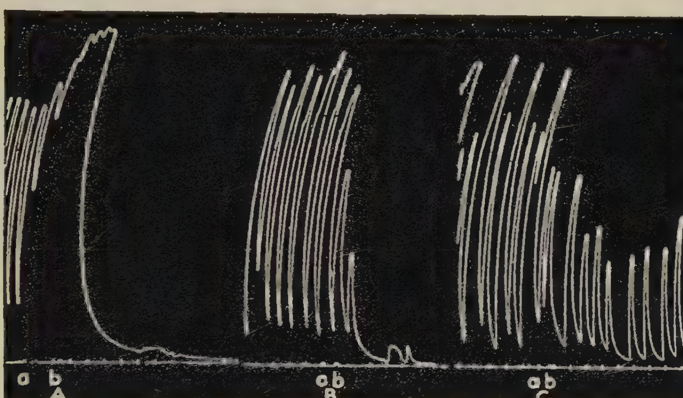


FIG. 11. 2. **Estimation of the Adrenaline Content of Blood.** Adrenaline inhibits the Rhythmic Contractions of the Intestinal Muscles of a Cat.

The muscle had been undergoing spontaneous contractions in Ringer's Solution (upstroke of the lever indicates contraction). At *a*, in each case, this solution was removed and at *b* was replaced by cat's blood containing: A, 1 : 1,000,000 adrenaline, B, 1 : 2,000,000 adrenaline, and C, 1 : 3,000,000 adrenaline. After a preliminary contraction, due to the action of the blood, inhibition of the spontaneous activity was produced, which was complete in the first two cases, and well marked in the last. (Cannon).

sympathetic system, but which do not abolish the contractile response of the muscle to direct electrical stimulation, or to the application of certain other drugs, such as barium chloride. These facts are usually summarised by the statement that adrenaline acts on an excitatory mechanism in the tissue, whereas certain abnormal stimuli (*e.g.* an electric shock and barium chloride) can act more directly on the contractile mechanism.

Small quantities of adrenaline in tissue extracts or body fluids can be estimated by matching their effects with those of a known concentration of the pure substance on isolated muscles suspended in Ringer's solution. The contraction of the rabbit's uterus, or relaxation of its intestine, are convenient indicators for this purpose and are sensitive to concentrations varying from 1 in 10^{10} to 1 in 10^{6}, the latter producing a nearly maximal effect (Fig. 11. 2). Even more sensitive tests are available

which will detect a few micrograms. Such a quantity of adrenaline antagonises, for example, a contraction of the rat's isolated uterus due to the action of acetylcholine ; a similar dose of noradrenaline raises the arterial pressure of an anæsthetised rat which has previously been treated with hexamethonium. These responses serve to emphasise the extreme physiological potency of these substances.

The adrenal medulla is under the control of the splanchnic nerve, stimulation of which is followed by changes in the organism which are in part due to direct stimulation of the sympathetic system, and in part due to the discharge of adrenaline, which reinforces and prolongs the immediate effects of the nervous stimulation (this is shown in Fig. 2. 5, p. 43). It has been estimated that the normal concentration

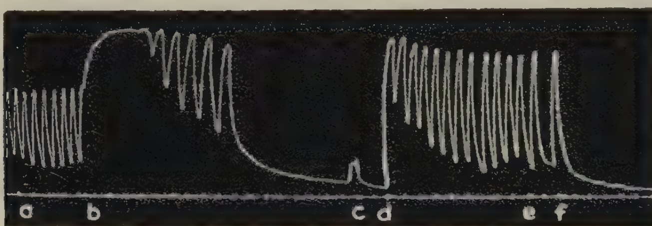


Fig. 11. 3. **The Secretion of Adrenaline during Excitement. Record of Rhythmically Contracting Intestinal Muscle.**

The muscle was initially beating in Ringer's solution. At *a*, this solution was removed, and at *b*, was replaced by blood taken from the inferior vena cava of a cat which had been excited by being " barked at by a dog for 15 minutes." After a preliminary contraction complete inhibition was produced. At *c*, this blood was removed, and at *d*, was replaced by blood obtained in a similar manner from a quiet cat ; the muscle immediately began to contract rhythmically. This blood was removed at *e*, and " excited " blood again added at *f* : the contractions ceased. Compare this with Fig. 11. 2, which shows the effect of blood containing known amounts of adrenaline on an intestinal muscle under similar conditions. (Cannon.)

of adrenaline in the blood leaving the adrenal glands is equivalent to a few hundredths of a microgram per minute per kilogram body weight under ordinary resting conditions. This is reduced to a negligible quantity after section of the splanchnic nerves and may be increased forty-fold by electrical stimulation of these nerves.

Reflex stimulation of the sympathetic system is produced under a variety of conditions, such as (1) physical exercise, (2) emotional states of fear and rage, (3) pain produced by stimulation of a sensory nerve, (4) asphyxia, (5) exposure to cold, and (6) general anæsthesia. The effects on the organs of the body are similar to those described above as resulting from injection of adrenaline or noradrenaline, and they may be summarised as a mobilisation of those resources of the body concerned in the response to an emergency and an inhibition of those functions which are not of immediate importance under such conditions. This is the foundation of the *emergency theory* of the function of the adrenal

medulla. It asserts that many of the physical changes which accompany intense emotion are due to the activity of the sympathetic and adrenal systems. Such activity results in a rapid pulse and rise in blood pressure, hair " standing on end," dilatation of the pupil and inhibition of the digestive processes, which are among the well-known signs of emotional excitement. The output of the heart is increased and the flow of blood directed primarily to the skeletal muscles, at the expense of that to the skin and abdominal organs. The crucial test of the emergency theory depends upon direct observation of the rate of discharge of adrenaline in the adrenal veins of an animal suitably stimulated. Exposing a cat to a barking dog, for example, increases the secretion of adrenaline, as shown in Fig. 11. 3. The emergency theory seems the more likely because it fits in happily with the conception of an organism well adapted to changes in its environment.

Tumours of the adrenal medulla may result in a rare, but characteristic clinical picture. Such chromaffin tumours are associated with attacks of tachycardia and high blood pressure. Extracts of such tumours have been shown to contain noradrenaline as well as adrenaline.

The Functions of the Adrenal Cortex

Removal or Disease of the Adrenal Glands. Removal of both adrenal glands is fatal within a period of one or two weeks. This result is due to loss of the adrenal cortex, for it has been clearly shown that removal of all adrenal medullary tissue, providing some cortical tissue is left *in situ*, is not followed by any severe abnormality. The completely adrenalectomised animal, however, after recovery from the shock of the operation, may appear normal for a few days. Then loss of appetite, muscular weakness, reduced temperature and blood pressure, diarrhœa, vomiting and anuria supervene. Loss of weight occurs and death follows usually within a few days of the first sign of illness. Addison's disease in man is a condition most commonly due to infection of the adrenal glands with *B. tuberculosis*. The clinical picture in the human closely resembles that produced by adrenalectomy in the experimental animal. It is characterised by muscular weakness, loss of weight, low blood pressure, low body temperature and vomiting. An additional feature, which has not been reproduced in animals, is bronzing of the skin and mucous membrane of the mouth in areas of pressure or friction. The disease usually has a gradual onset and runs a prolonged course during which acute crises of adrenal insufficiency occur from time to time.

In both experimental animals and man, loss of adrenal cortical tissue results in an increased excretion of sodium, and thus of water, by the kidney and a diminished excretion of potassium. As a result of this change in urinary excretion, concentration of the blood occurs which is probably responsible for the low blood pressure and renal failure and therefore for the raised concentrations of urea, phosphate and creatinine in the blood. In conformity with this it is found that administration of large quantities of sodium chloride causes a dramatic improvement in

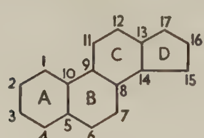
the condition of deficient men and animals. The concentrations of sodium and potassium in the blood are restored to normal with corresponding improvement in the blood volume, blood pressure and renal function. Such treatment may prolong life indefinitely. Certain abnormalities due to adrenal cortical deficiency are, however, not relieved by sodium chloride administration. These are : (i) disturbances in carbohydrate metabolism resulting in a tendency to hypoglycæmia ; (ii) muscular weakness as demonstrated experimentally by the rapid onset of fatigue to repeated electrical stimulation of the muscle ; (iii) disturbances in lymphoid tissue ; and (iv) a very low resistance to conditions of stress.

Adrenal Cortical Hormone. It has been firmly established that the effects of adrenalectomy are due to the loss of a hormone or hormones. Nearly thirty steroids have been isolated from extracts of adrenal cortical tissue, and of these some seven have been found to be active in maintaining the life of an animal lacking adrenal cortical tissue.

The physiologically active adrenal steroids may be divided into three groups.

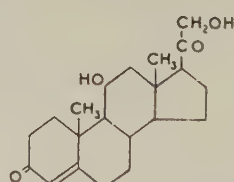
(1) Steroids with an oxygen or hydroxyl group at the C 11 position (Fig. 11. 4). These compounds correct the abnormalities of carbohydrate metabolism, of the performance of muscular work, of the thymus and lymphoid tissue and of the blood cells of adrenalectomised animals. They may be referred to as the gluco-corticoids.

(2) Steroids lacking an oxygen or hydroxyl group at the C 11 position, such as deoxycorticosterone. These compounds are particularly active in maintaining the life and growth of adrenalectomised animals and in correcting the disturbances of water and electrolyte metabolism, and are therefore referred to as the mineralo-corticoids.

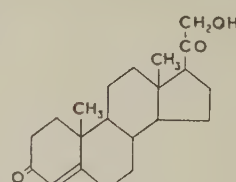


Cyclo-penteno-perhydro-phenanthrene

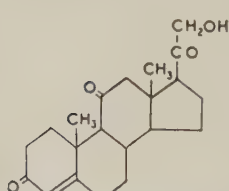
The Steroid Ring Structure



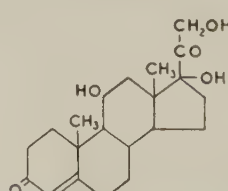
Corticosterone (Kendall's Compound B)



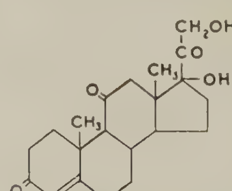
11 deoxycorticosterone (as the acetate = DOCA)



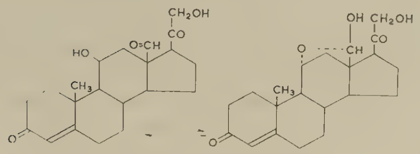
11 dehydrocorticosterone (Kendall's Compound A)



17 hydroxycorticosterone (Kendall's Compound F)



11 dehydro 17 hydroxy-corticosterone (Kendall's Compound E) Cortisone



Aldosterone (exists in two forms).

Fig. 11. 4. Structural Formulæ of Some Compounds related to the Adrenal Cortical Hormones.

(3) Steroids related to the sex hormones, including progesterone (compare Fig. 11. 4 with Fig. 10. 6 (p. 285)).

Studies of the steroids present in the blood of the adrenal vein have demonstrated the presence of compound B (mainly) in the rabbit, and compound F (mainly) in the monkey and man. Other forms examined were found to have varying ratios of compound B to compound F present. However, the hormone pattern present in adrenal vein blood was found to remain constant in any one species under varying degrees of activity of the adrenal cortex. Deoxycorticosterone has not been found present in significant amounts in either adrenal cortical extracts or in adrenal vein blood. A steroid which is relatively unstable but highly active in affecting the electrolytic balance of the body has, however, been detected in adrenal vein blood ; it has been crystallised, and its chemical structure is known. It has been given the name " aldosterone " since it possesses an aldehyde group at C-18. Studies of the above type have also shown that the rate of secretion of adrenal cortical hormone is normally very high if compared with the amount of hormone stored in the gland at any one time. The adrenal cortex has, therefore, a high rate of turnover but a low capacity for storage.

The actions of the adrenal cortical steroids can best be studied by administration of these substances to adrenalectomised animals. Administration of large doses by mouth or small doses by injection will maintain the life and growth of such animals. The carbohydrate metabolism is restored to normal by depression of the glucose utilisation by the tissues, an increased formation of glucose from protein and an increased glycogen storage. The electrolyte and water balance of the body is also restored to normal since the sodium and water excretion is reduced and the potassium excretion increased. The rapid onset of fatigue in the muscles is abolished. A striking effect of the glucocorticoids is the atrophy of the thymus and lymphoid tissue and the lymphopenia and eosinopenia which they produce. As mentioned above, the adrenalectomised animal is extremely sensitive to stress. Exposure of the normal animal to relatively mild stress stimuli (the prick of a hypodermic needle or emotional excitement), as well as to more severe stresses (extremes of environmental temperature, radiations, starvation, hypoxia, hypoglycæmia, infections, surgical trauma, muscular exercise, or administration of drugs, toxins or anæsthetics and so on), will result in greatly increased activity of the adrenal cortex. The adrenalectomised animal, however, does not adapt normally to

stress and will succumb to degrees of stress that may be well tolerated by the normal animal. It is worth emphasising that both the adrenal medulla and adrenal cortex are involved in the reaction of the body to conditions of emergency or stress, but that the severe and often fatal nature of stresses in the adrenalectomised animal is due to the loss of the adrenal cortex. Administration of relatively large doses of gluco-corticoids to the adrenalectomised animal will restore nearly normal resistance to stress. The mineralocorticoids are ineffective in this respect.

The adrenal cortical steroids or their metabolic products are excreted in the urine in two forms, as neutral 17-ketosteroids and as urinary corticoids. In the male the neutral 17-ketosteroids are also partly derived from the testis hormone. Clinically it has been found that disease or atrophy of the adrenal cortex (Addison's disease and Simmond's disease) is associated with excretion of decreased amount of these end products, whilst hyperfunctional states of the adrenal cortex are associated with excretion of increased amounts.

Control of the Adrenal Cortex. It is likely that the mineralo-corticoid, aldosterone, is secreted by the zona glomerulosa of the adrenal cortex, and that the gluco-corticoids, corticosterone and 17-hydroxycorticosterone, are secreted by the zona fasciculata. The main factor regulating the release of the gluco-corticoids is the secretion of the adreno-corticotrophic hormone (often shortened to A.C.T.H.) by the anterior pituitary gland. Following hypophysectomy, or disease of the pituitary gland, the adrenal cortex atrophies and the production of gluco-corticoids largely ceases. Following injurious stimuli or stresses of an emotional or physical type, the secretion of the adrenocorticotrophic hormone is rapidly and markedly increased with a corresponding rise in discharge of the gluco-corticoids from the adrenal gland. On the other hand the secretion of aldosterone seems largely, though not completely, independent of pituitary control. It is, however, influenced by changes in the electrolyte or water content of the body. In patients with œdema, due to some abnormality of the liver, kidneys or heart, the secretion of aldosterone is increased.

Hyperfunctional States of the Adrenal Cortex. It is of interest that the abnormalities observed after adrenalectomy or in Addison's disease are due to the loss of gluco-corticoids and mineralo-corticoids, whilst the main abnormality seen in hyperfunctional states of the adrenal cortex seem due to excessive secretion of the androgenic steroids. The hyperfunctional state, the so-called adrenogenital syndrome, is associated with a hyperplastic adrenal cortex, or with a benign or malignant tumour of the adrenal cortex.

In females of any age, from intrauterine life onwards, overgrowth of cortical tissue may give rise to a form of hermaphroditism. Hair appears on the face, the pubic hair assumes a male distribution, the voice becomes deeper, the clitoris enlarges and amenorrhœa supervenes. In boys, precocious sexual maturity may take place and is associated with muscular hypertrophy. In both sexes, mental changes appropriate

to these conditions may occompany them, *i.e.* a masculine outlook in women and girls, and a mature attraction for the opposite sex in boys.

A condition known as primary aldosteronism has also been described. It is due to an adrenal cortical tumour, secreting excess amounts of aldosterone.

THE THYROID GLAND

The thyroid gland consists of two lobes connected by an isthmus, and is situated anterior and lateral to the upper part of the trachea and larynx.

Removal of the Thyroid and Myxœdema. Enlargements of the thyroid gland (goitre) have been known for a long time. Such enlarged glands may press on neighbouring structures and in such circumstances the entire gland was at one time surgically removed. Some months after removal a condition called " *cachexia strumipriva* " developed, and its similarity to that termed *myxœdema*, known to occur spontaneously in middle-aged or elderly adults, was noticed. Both states are now known to be due to lack of thyroid hormone (as a result of either removal or disease of the thyroid), as is the allied condition *cretinism*, which occurs in young children.

Myxœdema is characterised by a thickening of the subcutaneous tissue, coarsening of the features, dryness and pallor of the skin and falling out of the hair. The patient is mentally dull, sluggish in action, with a slow pulse and perhaps a subnormal temperature. A history of constipation and, in women, amenorrhœa may be obtained. A striking feature is the reduction in basal metabolic rate, which may be 30 or 40 per cent. below the normal (Fig. 11. 5). Nitrogen excretion is diminished and the serum contains abnormally large quantities of cholesterol.

Cretinism occurs fairly commonly in children born in geographic regions where goitre is common. In its most severe form it begins to be apparent at the age of three months, but in addition to the features it has in common with myxœdema, others occur due to *arrested development.* The child, even should he grow up, never achieves a mental capacity greater than that of a normal child of about three years, and may actually lack initiative to feed himself at the age of twelve years. The long bones fail to grow at the epiphyses, leading to much loss of stature, but the skull does enlarge, giving a characteristic deformity. The fontanelles are late in closing. The visceral organs also grow, so the individual becomes pot-bellied. The muscles become disproportionately large (without achieving any great strength) with the consequence that the tongue is too big for the mouth, from which it protrudes, causing a dribble of saliva. To these grotesque abnormalities may be added pads of fat, deposited on the shoulders and buttocks.

Experimental removal of the thyroid in animals produces conditions similar to those seen in the human. The marked diminution in metabolic rate, slow pulse, lowered body temperature, and the disturbances in general bodily, sexual and mental development may be clearly observed.

The effect of the thyroid gland on bodily development is striking in amphibians. In these forms removal of the thyroid of the tadpole results in greatly delayed metamorphosis so that enlarged tadpoles are produced instead of normal-sized, small frogs.

Administration of thyroid tissue or thyroid extracts by mouth will, if begun early enough, prevent development of the abnormalities of the

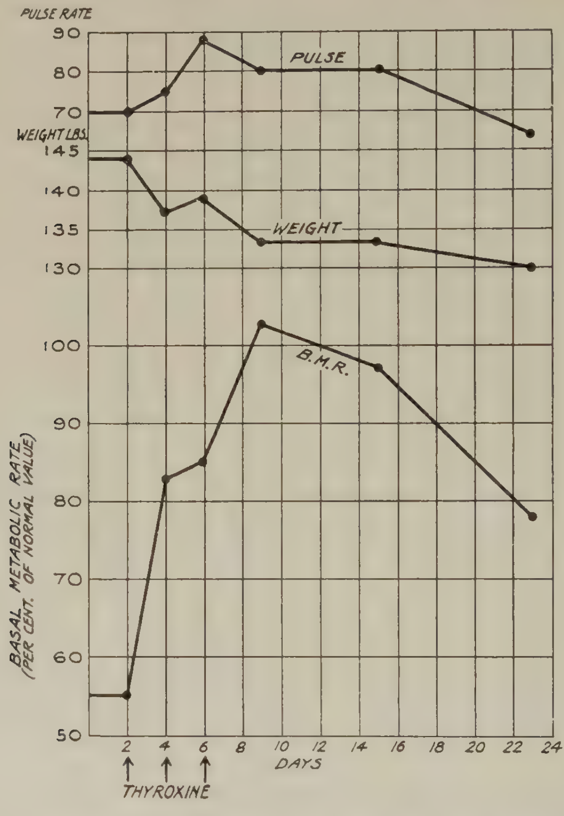


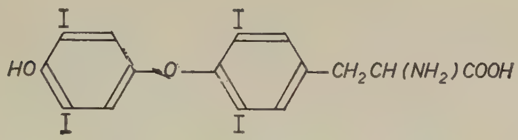
FIG. 11. 5. The Action of **Thyroxine** on the Basal Metabolic Rate, the Weight and the Pulse Frequency of a Patient with Myxœdema.

5 mg. of synthetic thyroxine was injected intravenously on the second day, 4 mg. on the fourth day, and 5 mg. on the sixth day ; in addition to the increase in basal metabolic rate, fall in weight and rise in pulse frequency shown on the chart, the injections produced a rise in temperature, diuresis and diarrhœa. (After Lyon.)

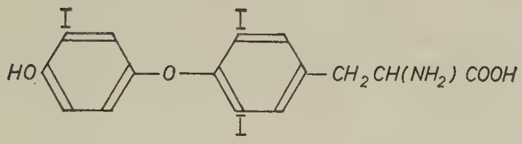
hypothyroid state. It is especially important that thyroid therapy should begin early in cases of cretinism, for after several years, irreversible changes will have occurred and improvement is all that can be hoped for. Cessation of treatment in either myxœdema or cretinism leads to a prompt relapse into thyroid deficiency (Fig. 11. 5).

Thyroid Hormone. The thyroid gland has a remarkable ability to extract inorganic iodine from the plasma, even though the concentration

of such iodine is many hundreds of times less in the plasma than in the gland. Inorganic iodide in the gland is rapidly converted to di-iodo-tyrosine and so to thyroxine and tri-iodothyronine, which are then stored in protein combination in the colloid of the vesicles as thyroglobulin.



Thyroxine (3 : 5 : 3′ : 5′ tetra-iodothyronine)



(3 : 5 : 3′ tri-iodothyronine)

It is clear that thyroglobulin is not the hormone secreted into the blood. L-thyroxine and L-tri-iodothyronine have been detected in plasma, and it is likely that the thyroglobulin undergoes hydrolysis in the gland to liberate these two compounds which circulate in the blood stream in loose linkage with the plasma proteins.

Both L-thyroxine and L-tri-iodothyronine are more active than their dextro-rotatory isomers. In the various laboratory and clinical tests that have been employed up to the present time L-tri-iodothyronine has been found to be approximately five times as active as L-thyroxine. L-tri-iodothyronine also acts more quickly than L-thyroxine.

Action of the Thyroid Hormone. Desiccated thyroid powder, thyroxine or tri-iodothyronine will effectively restore thyroidectomised animals or patients suffering from myxœdema to a normal state. Administration of excessive amounts of thyroid hormone results in a much increased metabolic rate, increased nitrogen excretion, loss of weight, rapid pulse, raised body temperature, sweating, flushing, a heightened sensitivity of the autonomic nervous system, fine muscular tremors and, in man, marked anxiety. The mechanism whereby the changes are brought about is not understood. An effect on tissue metabolism may be primary to some of the above effects, but it should be remembered that di-nitrophenol, a drug that raises the metabolic rate, does not produce the other effects characteristic of overdosage with thyroid hormone.

Thyroid Activity and its Control. Until a few years ago the easiest way to determine the activity of the thyroid gland was to measure the metabolic rate or oxygen consumption of an animal. Since the introduction of radioactive isotopes it has become possible to measure thyroid activity more quickly and more directly by making use of radioactive iodine, ^{131}I. Injection of a chemically minute, and metabolically negligible, amount of ^{131}I into an animal makes it possible to follow the

rate of uptake of this radio-iodine by the thyroid, and also to follow it as it passes through the various phases in the gland to be incorporated as part of the molecule of thyroglobulin or thyroid hormone. A large proportion of injected ^{131}I is excreted through the kidneys, which is a factor that must be remembered when using the thyroid uptake of radio-iodine as a measure of thyroid activity. An alternative method of measuring thyroid activity in laboratory animals is to measure the rate of release of radioactive hormone from the gland. Forty-eight hours after the administration of a small dose of ^{131}I the radio-iodine left in the body is nearly all in the form of radioactive hormone in the thyroid gland. Measurement of the rate at which the radioactivity in the gland declines gives values which are, under standardised conditions, proportional to the rate of discharge of thyroid hormone. It is important to emphasise that the thyroid metabolises radioactive iodine in the same way as non-radioactive iodine, ^{127}I, and that the methods available for detecting and measuring ^{131}I are so sensitive that the amounts administered to perform the above tests do not in any way disturb iodine metabolism.

Thyroid activity is regulated by the secretion of the thyrotrophic hormone (T.S.H.) by the anterior pituitary gland. Hypophysectomy greatly reduces thyroid activity although such activity is not entirely abolished. The changes in the external environment which lead to alterations in the rate of secretion of thyroid hormone appear to do so by affecting the rate of secretion of T.S.H. Exposure to a cold environment exerts a pronounced stimulating effect on the thyroid. Other environmental stimuli, which may be interpreted as emotional or physical stresses (restraint, painful stimuli, operative procedures, hæmorrhage and so on) have been reported as exerting pronounced inhibitory effects on the thyroid gland. The thyroid gland of the female shows increased activity at various stages of the reproductive cycle, that is at puberty and during pregnancy, but the mechanism of these changes is uncertain, although they can probably be correlated with changes in ovarian activity.

Goitre. The term goitre refers to any visible swelling of the thyroid gland. A goitrous gland may be associated with a state of hypo-, eu-, or hyper-thyroidism. One common form of goitre, **endemic goitre,** is mainly due to a lack of iodine in the diet and tends to occur in various geographical areas in which the iodine content of the water or of the soil (and therefore in the vegetable food of the region) is low. This type of goitre occurs, for example, in the British Isles in a belt running roughly from Somerset through Derbyshire to Durham, but is not restricted to this area. A goitrous population of this kind shows also a high incidence of cretinism and goitre in the children. Addition of small quantities of iodine to the diet, usually by iodising the table salt consumed, leads to a reduction in the size of goitrous glands and tends to reduce to negligible proportions the incidence of goitre and cretinism in the children.

Exophthalmic Goitre (Graves' or Basedow's Disease). This condition of hyperthyroidism, first described by Parry in 1825, is associated with a

diffuse hyperplasia of the thyroid gland. *The basal metabolic rate is greatly increased* and the skin is flushed and moist. There is loss of weight and tachycardia, the apex beat being diffuse, and its position suggesting an enlargement of the heart which is not, however, apparent post-mortem. Mental activity is accelerated, with exaggerated responses to sensory and emotional stimuli. Just as the increased frequency of the heart beat may cause cardiac failure, so the heightened susceptibility of the nervous system may lead to insanity. Protrusion of the eyeballs, for which there is no complete explanation (though it can be produced in animals by administration of the thyrotrophic hormone from the anterior pituitary gland), is typical. It is uncertain whether the over-activity of the thyroid cells is a primary defect, or is secondary to some other change, such as excess of the thyrotrophic hormone.

Antithyroid Substances. Many substances are known to inhibit the

FIG. 11. 6. **Exophthalmic goitre.**
Note the swelling in the neck (goitre) due to enlargement of the thyroid gland and the protruding eye-balls (exophthalmos). (Parsons.)

uptake of iodine by the thyroid, or to inhibit the organic-binding of iodine and so prevent the formation of di-iodotyrosine and thyroxine.

(*a*) Iodine uptake from the blood may be reduced by administration of *excess* iodine (the reason for this is not known), by thiocyanates and by various common foodstuffs, such as cabbage.

(*b*) Incorporation of iodine in the gland into organic combination may be blocked by the aminobenzene derivatives including the sulphonamides, and by the thioureas and thiouracils. One of the most potent of these substances is propyl-thiouracil. Experimentally it has been found that administration of these compounds is particularly effective in producing a goitrous enlargement of the thyroid. The probable sequence of events is that the antithyroid compound inhibits the synthesis of thyroid hormone, the concentration of the hormone in the blood falls, and this results in increased secretion of thyrotrophic hormone and thereby hyperplasia of the thyroid.

THE PARATHYROID GLANDS

Two pairs of very vascular bodies, about the size of lentils, attracted the attention of surgeons in the early days of therapeutic removal of the

thyroid glands. Unless one or more of them was left undisturbed, tetany supervened and the consequences were usually fatal.

The parathyroid glands are necessary for life. After sudden extirpation of all parathyroid tissue, there are no untoward effects for the first three or four days ; but then follow periodic spasms of the musculature, high fever, increased irritability of the peripheral nerves and death due to asphyxia during one of the " tetanic " convulsions. This condition of *tetany* is associated with an abnormally low concentration of calcium

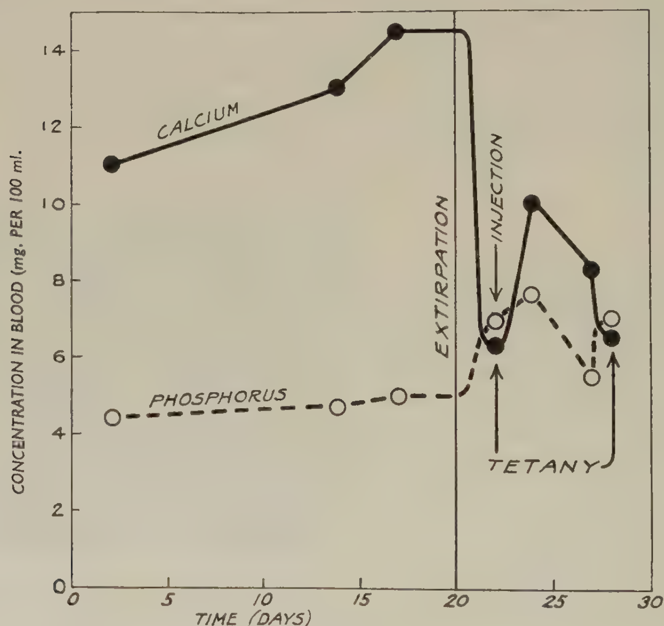


FIG. 11. 7 The Effect of **Extirpation of the Parathyroid Glands and Subsequent Injections of Parathormone** on the Concentrations of Calcium and Inorganic Phosphorus in the Blood, showing their actions in inducing and preventing Tetany in a Dog.

Tetany occurred whenever the calcium concentration fell below about 7 mg. per 100 ml. (After Weaver and Reed.)

salts in the blood ; it can be alleviated by administration (oral or intravenous) of calcium salts, and more completely controlled by periodic injections of parathyroid extract.

Deprivation of parathyroid by surgical removal of the gland is always followed by a fall in the concentration of calcium in the blood, which, if sufficient, is associated with tetany. Injection of parathyroid extract (parathormone), on the other hand, results in an increased concentration of calcium in the blood, the extra calcium being derived from the bones. These relations are illustrated in Fig. 11. 7, which also indicates the characteristic variations in concentration of inorganic

phosphate produced by the same agents ; the phosphate and the calcium concentrations vary in opposite directions.

The parathyroid glands are, therefore, related to the metabolism of calcium salts in the body, and the effects produced by excessive and inadequate functioning of these glands are essentially the same as those produced by excess and deficiency of calcium.

The normal concentration of calcium salts is about 10 mg. Ca per 100 ml. of plasma ; about half of the calcium is held by protein and is indiffusible through collodion membranes.

Hypercalcœmia (more than 20 mg. Ca per 100 ml. plasma) when produced in the dog by repeated large doses of the hormone, is associated with a rise in the viscosity and the osmotic pressure of the blood and with an increase in the concentration of phosphates, proteins and urea ; diarrhœa, vomiting, vascular failure, coma and death may ensue. Parathormone injections have been given to man in cases of lead poisoning to remove the lead which has accumulated in the bones. Calcium comes away with the lead and produces a hypercalcæmia, and cases are recorded in which the blood calcium rose to nearly 20 mg. per 100 ml. without producing severe symptoms.

Hypocalcœmia (less than 7 mg. Ca per 100 ml. plasma) results in the muscular twitchings and tetanic convulsions mentioned above ; the effect of immersing isolated nerves in saline solutions deficient in calcium may be recalled in this connection (Chapter 12, p. 371).

Ancillary evidence connecting the parathyroid glands and calcium metabolism is derived from the pathological changes in the glands associated with diseases involving failure in the normal calcification of bone. For example, overgrowth of the glands accompanies malnutrition on calcium-deficient diets, rickets, and osteomalacia ; a similar hypertrophy is usual during pregnancy and lactation. Partial excision of the glands, moreover, results in deposition of defective dentine in the teeth, and in an imperfect development of callus about bone fractures—abnormalities which can be rectified, at least in rats, by implanting additional parathyroid tissue.

Very large quantities of vitamin D have been found to raise the concentration of calcium in the serum (about 100 times the anti-rachitic dose is needed). This is apparently the result of a direct action in releasing calcium from the bones and of promoting absorption from the gut, as already mentioned in Chapter 7. The parathyroid gland is not involved. It has been found, also, that a substance known as tachysterol, which is an intermediate product in the formation of calciferol (vitamin D_2) by irradiation of ergosterol, is also effective in a similar way, and that di-hydroxy-tachysterol is even more effective. This substance, known as " A.T.10," is in use to raise the serum calcium concentration, and to counteract the effects of chronic hypoparathyroidism.

Action of Parathyroid Hormone. It is probable that parathyroid hormone (parathormone) exerts a double action. It seems to act upon the cells of the renal tubules to decrease the reabsorption of phosphate from the tubular fluid. As a consequence the concentration of inorganic

phosphate in the blood falls, calcium phosphate is mobilised from bone and the blood calcium concentration therefore rises. If this were the only action of parathyroid hormone, then it would lose activity if the kidneys were removed. However, experiments designed to test this possibility indicate that parathormone is still active in the nephrectomised animal. It is likely that it also exerts a direct action on bone in mobilising calcium.

Control of Parathyroid Secretion. Since parathyroid transplants appear capable of maintaining normal function it is unlikely that the glands are regulated *viâ* a secretomotor nerve supply. There is also little reason to believe that the parathyroids are controlled by an anterior pituitary hormone. Rather the evidence indicates that the activity of the gland is influenced directly by the calcium or phosphate concentration of the blood, for hyperplasia of the parathyroids appears to be associated with conditions in which the blood calcium concentration is low (rickets, chronic nephritis), and conversely the parathyroids are reduced in size if the blood calcium concentration is raised by varying the diet.

THE PITUITARY GLAND

The pituitary gland weighs about ½ gram in man, and is attached to a region of the base of the brain, the hypothalamus, by the *pituitary stalk*. It is a well-protected structure lying in a fossa in the sphenoid bone known as the *sella turcica*. The gland is composed of two parts, the **neurohypophysis** which is embryologically an outgrowth from the hypothalamus, and the **adenohypophysis**, derived from an evagination from the roof of the embryonic mouth, Rathke's pouch. This pouch loses all connection with the mouth, but still retains its original lumen as a cleft in the fully developed organ. Since the adult gland is easily split by tearing into two portions, and since the line of cleavage occurs through this cleft, the two parts so obtained are commonly referred to as the *anterior* and *posterior* lobes. It should be remembered, however, that the term posterior lobe also includes the pars intermedia and therefore is not a clear definition of a functional region of the gland. Nevertheless the term *posterior lobe extract* is still in common use. The generally accepted terminology of the various parts of the pituitary is shown in Fig. 11. 8. The adenohypophysis is subdivided into the pars distalis (the main secretory part), the pars tuberalis which is a leaflike extension forming a collar around the upper end of the pituitary stalk, and the pars intermedia. The neurohypophysis is subdivided into the median eminence, the infundibular stem and the infundibular process.

Nerve Supply and Blood Supply of the Pituitary. The main nerve supply of the pituitary is derived from the hypothalamus. The supraoptic and paraventricular nuclei send a very rich bundle of unmyelinated nerve fibres, the supraoptico-hypophysial tract, to end in all three parts of the neurohypophysis (Fig. 11. 8). In contrast to this rich innerva-

tion of the neurohypophysis, the anterior lobe receives a very scanty nerve supply, which probably consists of sympathetic vasomotor nerves.

The blood supply of the pituitary gland (Fig. 11. 9) is derived directly from the internal carotid artery and the circle of Willis. The venous drainage passes into nearby dural venous sinuses. The pars distalis possesses in addition a second blood supply in the form of a portal circulation, which is arranged as follows. Arterial twigs supply a rich vascular plexus in the pars tuberalis, and from this plexus spring a wealth of capillary loops and sinusoids (the primary plexus) that

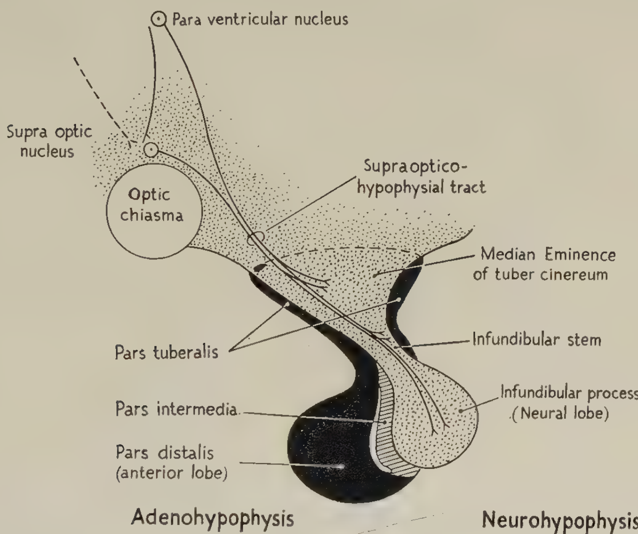


FIG. 11. 8. Diagram of a sagittal section through the hypothalamus and pituitary gland to show the various subdivisions of the adenohypophysis and neurohypophysis. The supraoptico-hypophysial tract is shown arising in the hypothalamus and terminating in all three parts of the neurohypophysis.

penetrate the neural tissue of the median eminence. From the primary plexus the blood is collected into the trunks of the portal vessels, which pass down the pituitary stalk, and are distributed into the sinusoids of the anterior lobe.

Hypophysectomy. Removal of the pituitary gland is a relatively simple operative procedure in most laboratory animals (rats, rabbits, ferrets, cats and others) when the parapharyngeal route is used and the gland is removed through a drill hole in the base of the skull. The operation is not immediately fatal. Under favourable conditions the life span of the hypophysectomised animal is about half that of the normal. However, the operation is followed by dramatic changes that are due mainly to the loss of the anterior lobe. Posterior lobe function may not be greatly disturbed since the median eminence and the infundibular stem, which form part of the neurohypophysis, are left *in situ*. It might

be argued that the pars tuberalis of the adenohypophysis is also left intact, but it is doubtful whether this part of the gland has any secretory function.

The effects of hypophysectomy may be summarised as follows. Growth (in the young animal) is greatly slowed or ceases entirely. The ovaries and reproductive tract fail to mature or undergo atrophy. The thyroid and adrenal cortex atrophy, though the adrenal medulla is unaffected. Increased sensitivity to insulin and hypoglycæmia occur, and in some animals may result in early post-operative death. If the operation is performed in the lactating female, lactation promptly ceases. Since the anterior pituitary is of such importance in main-

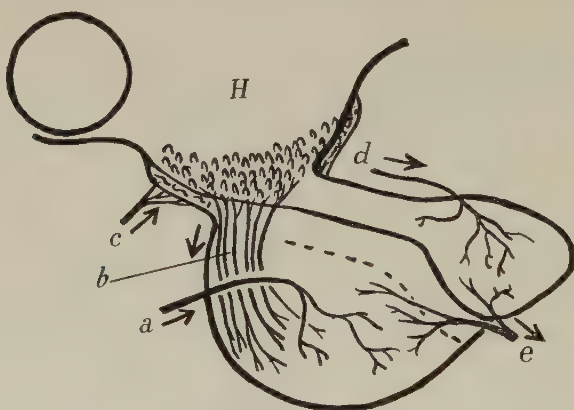


Fig. 11. 9. Diagram of a sagittal section through the hypothalamus (H) and pituitary gland of a rabbit to illustrate the vascular supply of the pituitary gland. The anterior lobe receives a systemic arterial supply (*a*) from the internal carotid artery, and a portal supply (*b*). The portal vessels originate in the primary plexus in the tuber cinereum which is itself fed by arterial twigs (*c*) from the carotid artery and the posterior communicating artery. The posterior lobe or infundibular process acquires a separate blood supply (*d*). The venous drainage (*e*) is into surrounding venous sinuses in the dura.

taining the activity of the ovary, testes, thyroid and adrenal cortex, these latter glands are sometimes referred to as pituitary " target organs." Although these target organs undergo marked atrophy in the absence of the pituitary, it is to be noted that some slight basal level of activity is maintained. For example, the adrenal cortex secretes sufficient hormone to maintain life, the thyroid secretes a small (about one-twentieth the normal) amount of thyroid hormone, the testis may still contain spermatogonia dividing into primary spermatocytes and the ovary shows the presence of primordial follicles developing to the stage of beginning antrum formation. This minimal level of activity seems to be inherent in the glands themselves and independent of external control.

Adenohypophysis

Hormones of Anterior Pituitary. Current views hold that at least six hormones are secreted by the anterior lobe—the gameto-kinetic or follicle-stimulating hormone (F.S.H.); luteinising or interstitial-cell-stimulating hormone (L.H. or I.C.S.H.); lactogenic or luteotrophic hormone (prolactin, Lt.H.); growth or somatotrophic hormone (S.T.H.); thyrotrophic hormone (T.S.H.) and adrenocorticotrophic hormone (A.C.T.H.). That the ill effects of hypophysectomy are due to hormonal deficiency may be shown by grafting anterior pituitary tissue from another animal beneath the hypothalamus of the completely hypophysectomised animal. Normal anterior pituitary function may be re-established following such a graft. A pituitary transplant placed elsewhere in the body, however, is not capable of maintaining normal function. This fact does not reflect on the endocrine status of the anterior pituitary, but rather indicates that the normal stimulus to activity of the gland is derived, in some way (see below), from the hypothalamus.

As a matter of terminology it is debatable whether the six hormones referred to above are hormones in the sense of being separately and independently secreted into the blood stream. They are all protein in nature and it is not yet possible to measure the quantities present in pituitary venous blood or to give figures representing their rate of secretion under different physiological states. For the time being they are, however, generally called hormones.

Gonadotrophic Hormones. (Follicle-stimulating hormone, F.S.H.; luteinising or interstitial-cell-stimulating hormone, L.H. or I.C.S.H.; and luteotrophic or lactogenic hormone, Lt.H. or prolactin.) F.S.H. and L.H. (or I.C.S.H.) are glycoproteins, the molecular weight of the latter being variously estimated at 40,000 or 100,000. Lt.H. is a simple protein with a molecular weight of about 30,000, and has been obtained in a crystalline form. The double actions of these hormones on the gonads have been discussed in Chapter 10, p. 294: F.S.H. and L.H. maintain the production of sperm and ova, and all three maintain the various internal secretions—androgens, œstrogens and progesterone. Lt.H. is necessary, also, for the *formation* of milk by the mammary glands (Chapter 10, p. 312).

Growth Hormone (Somatotrophic Hormone, S.T.H.). The growth hormone is a simple protein with a molecular weight of about 45,000. It has been obtained in a pure crystalline form. This hormone will correct the arrested growth of the hypophysectomised young animal, and, if given in excess, will lead to experimental gigantism (Fig. 11. 10). The hypoglycæmia and increased sensitivity to insulin of the hypophysectomised animal are probably due to loss of the growth hormone. Repeated injections of large amounts of growth hormone into adult dogs results in hyperglycæmia and glycosuria, and may lead to a state of permanent diabetes; while in younger animals, in which the islet tissue of the pancreas is able to compensate with increased insulin

secretion, protein is laid down and increased growth ensues. What used to be called the diabetogenic hormone is now known to be identical with the growth hormone. The activity of the pituitary gland in carbohydrate metabolism has been discussed in Chapter 6 (p. 198).

Growth hormone preparations are species specific. Extracts derived from ox and pig tissue were known to be disappointing in their effects on the primate. Now it has been established that extracts of primate pituitary tissue are active in causing nitrogen retention, growth and typical effects on carbohydrate metabolism in primates.

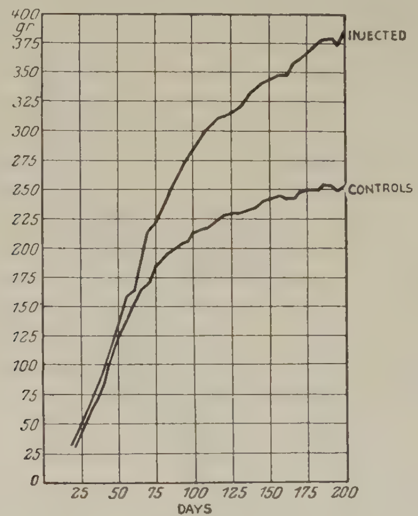


FIG. 11. 10. The Effect of Injection of Extracts of the **Anterior Lobe of the Pituitary Gland** on the Growth of Rats.

Two groups of thirty-eight rats each were selected ; the rats in one group served as controls, while those in the other were injected daily, and the average weight of the animals in each group was plotted against the time. Note the "gigantism" produced in the injected group. (Trendelenburg, after Evans and Long.)

Thyrotrophic Hormone (T.S.H.). This hormone has not been completely purified, but appears to be a glycoprotein with a molecular weight of about 10,000. Extirpation of the pars anterior of the pituitary body results in a reduction of the basal metabolic rate by some 30 per cent., and diminished nitrogen excretion. Conversely, if the thyrotrophic hormone is injected daily into hypophysectomised rats the basal metabolic rate is raised to normal in three or four days. That it acts through the thyroid may be shown by the discharge of colloid, and hyperplasia of the epithelial cells (a direct action of T.S.H. on peripheral tissues may be the exophthalmos which follows T.S.H., though not thyroxine, administration). When T.S.H. is injected into the normal animal, the metabolic rate is raised, but in spite of continued injection this returns to normal in two or three weeks and indeed falls to 20 or 30 per cent.

below normal if the injection is long continued. This is due to the gradual production of a substance which is antagonistic to the hormone. In view of the action of the thyroid extract on growth and development already described, it is necessary to state here that the lack of growth following hypophysectomy is not corrected by administration of thyroid gland, *i.e.* it is not attributable wholly to absence of the thyrotrophic hormone, though this deficiency plays its part.

Adrenocorticotrophic Hormone. A.C.T.H. is a simple protein of molecular weight 20,000. Hydrolysis yields a polypeptide of molecular weight about 1,200 which retains most of the biological activity of the protein. The atrophy of the adrenal cortex which follows hypophysectomy may be corrected by A.C.T.H. administration and excessive amounts of this hormone may lead to hypertrophy of the adrenal cortex. The action of A.C.T.H. on the adrenal cortex may also be observed by measuring the fall in cholesterol or ascorbic acid content of the adrenal gland which follows administration of this hormone. The ascorbic acid depletion test has been developed as a highly sensitive method of assay for A.C.T.H. Injection of A.C.T.H. in the normal animal causes increased release of the gluco-corticoids, with their corresponding actions on carbohydrate metabolism, lymphoid tissue, blood cells and response to stress. A.C.T.H. has therefore been used therapeutically, instead of cortisone, in the treatment of rheumatoid arthritis.

The Control of Anterior Pituitary Secretion. Endocrine activity in general is regulated mainly by the hypothalamus. This controls the neurohypophysis and adrenal medulla by a secretomotor nerve supply, and through its control on the secretion of the anterior pituitary trophic hormones also regulates, indirectly, the activity of the gonads, thyroid and adrenal cortex (Fig. 11. 11).

Under quiescent optimal conditions the resting rate of secretion of the pituitary trophic hormones seems to be set by the concentration in the circulating blood of the target-organ hormones. The gonadal, thyroid and adrenal cortical hormones exert an inhibitory influence over the secretion of their respective pituitary trophic hormone. It is uncertain whether this action is a direct one on the cells of the anterior pituitary ; it seems possible that the target-organ hormones affect some neural mechanism in the hypothalamus which in turn adjusts pituitary secretion. This feed-back mechanism is clearly of importance in setting a constant " base-line " activity of the glands.

Exteroceptive stimuli exert a profound influence on anterior pituitary secretion. Changing conditions of environmental light are particularly potent in many kinds of animal in varying the rate of F.S.H. secretion ; coitus in a number of mammals acts as a trigger stimulus to L.H. release. The tactile stimulus to the nipple of suckling appears to augment lactogenic secretion, whilst stressful or noxious stimuli will markedly increase A.C.T.H., and decrease T.S.H. secretion. These and other environmental stimuli probably act *viâ* nervous reflex paths to the hypothalamus which, in turn, affects the anterior pituitary.

The evidence indicates that exteroceptive stimuli are over-riding factors in regulating pituitary secretion and that they take precedence over the feed-back mechanism of the target organ hormones described above. For example, injection of cortisone will normally inhibit the secretion of A.C.T.H., but fails to do so in the presence of severe stress.

The anatomical path by which the hypothalamus influences the anterior pituitary has been much discussed. This gland lacks a rich

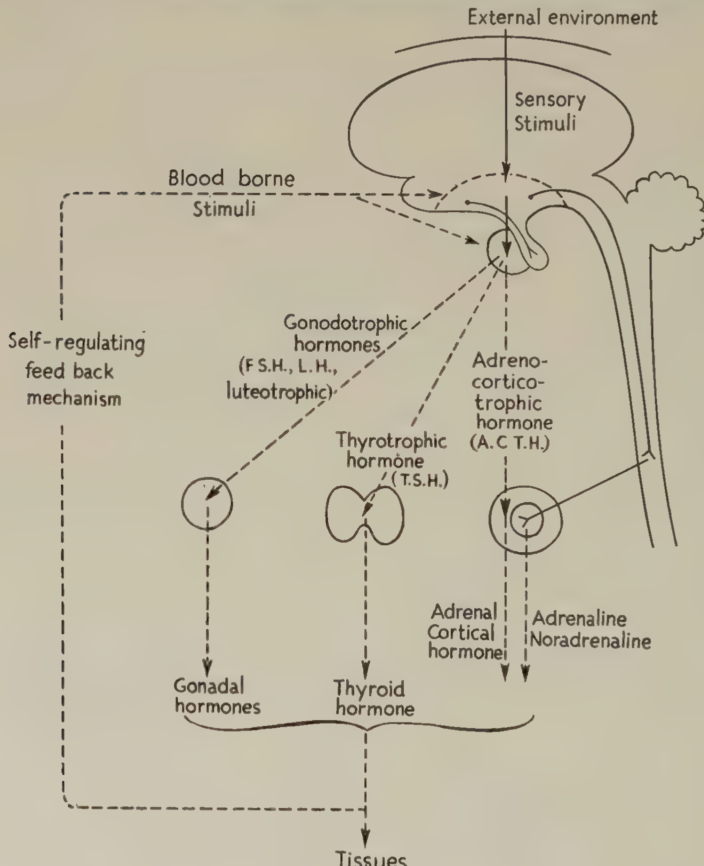


FIG. 11. 11. Diagram to illustrate neuro-endocrine inter-relations.

nerve supply, but it has been established that the pars distalis of the pituitary is connected to the tuber cinereum in all animals from amphibians to man by the hypophysial portal vessels, described on p. 335 above. The direction of blood flow in these vessels is *from* the tuber cinereum *to* the pituitary gland, and the general arrangement would seem well suited for the transport of substances formed in the tuber cinereum to the anterior lobe. It now seems likely that the hypothalamus and associated hypophysial vessels form a functional unit with the anterior lobe for reasons such as the following :

(*a*) Simple section of the pituitary stalk in the region of the trunks of the portal vessels is followed by variable results so far as the function of the anterior pituitary is concerned. In those animals in which normal anterior pituitary activity is regained, regeneration of the vascular connections across the site of the cut can be shown to have taken place. In those animals in which regeneration does not occur, or is prevented by the insertion of a plate, anterior pituitary function is very much reduced or abolished after operation, and remains so.

(*b*) A pituitary graft placed beneath the median eminence of the hypophysectomised animal obtains a blood supply from the primary plexus of the portal vessels and will maintain normal anterior pituitary function, whereas grafts placed beneath the temporal lobe or in other sites remote from the sella turcica may obtain an equally good blood supply but show very little, if any, sign of functional activity.

(*c*) Lesions of the hypothalamus may interfere with the release of gonadotrophic, thyrotrophic and adrenocorticotrophic hormones. Stimulation of the hypothalamus electrically has been shown to result in liberation of gonadotrophic and adrenocorticotrophic hormone. Similar stimuli applied directly to the anterior pituitary gland are ineffective in causing hormone release. This latter finding has been taken to indicate humoral (vascular), rather than direct nervous, control of the anterior lobe.

There can be little doubt that the hypothalamus and hypophysial portal vessels are intimately concerned with maintaining and regulating anterior pituitary activity, but the detailed mechanism is uncertain. On the present data the most likely hypothesis is that hypothalamic nerve fibres liberate some chemical transmitter(s) into the primary plexus which is carried by the vessels to excite or inhibit the activity of the anterior pituitary gland.

The hypothalamus seems to regulate not only the level of anterior pituitary activity, but also the hormonal pattern secreted. As mentioned above the pituitary gland of the male secretes little L.H., but if a male gland is transplanted into a hypophysectomised female in such a way that it becomes revascularised by the portal vessels it may support normal female sex cycles and pregnancy. Similarly it has been shown that anterior pituitary tissue from new-born animals is capable of supporting adult female sexual functions. It would appear then that anterior pituitary tissue is plastic in nature, and that the mosaic of hormones secreted is normally " set " by the central nervous system.

Clinical Hyperpituitary Syndromes. There are at least three types of cell in the anterior pituitary, chromaphobe (non-secreting mother cells), acidophile or α-cells and basophile or β-cells. Three types of tumour, or overgrowth (adenomata), are known corresponding to these cells.

Chromaphobe Adenoma. This is the commonest type of anterior pituitary lesion. Since the cells forming the tumour are non-secreting no signs of hyperpituitarism appear. The pressure of the tumour on surrounding normal secreting cells may, however, give rise to signs of hypopituitarism, and may also involve the hypothalamus.

Acidophile Adenoma. Since the acidophile cells are associated with secretion of the growth hormone, a secreting tumour of this cell type gives rise to *gigantism* if it arises before growth ceases, or *acromegaly* if it arises later in life. This latter condition is associated with overgrowth of the extremities and head, kyphosis, weakness, sexual hypofunction, and there may be visual disturbances from pressure of the tumour on the optic chiasma. There is frequently a decreased sugar tolerance and even hyperglycæmia or diabetes. Some types of dwarfism, on the other hand, may be due to deficiency of growth hormone. A hereditary lack of acidophile cells, for example, has been demonstrated in the pituitary gland of a strain of dwarf mice.

Basophile Adenoma. This tumour is associated with a condition known as Cushing's disease or pituitary basophilism, in which a variety of signs and symptoms may be exhibited. Among the more constant are adiposity, confined to the face, neck and trunk, often with tenderness ; kyphosis, possibly accompanied by rarefaction of the vertebræ ; sexual dystrophy which may start in females with precocious development, but which ends with amenorrhœa in women and impotence in men ; a tendency to a masculine distribution of hair in females ; a high coloured or dusky skin ; purple abdominal striæ, initiated by the stretching of the skin by abdominal fat ; high blood pressure and polycythæmia. The subjects are generally young adults and the suprarenal cortex is hypertrophied.

Neurohypophysis

Extracts of the neurohypophysis, commonly called posterior pituitary extracts, possess a variety of pharmacological activities, but only some of these represent normal physiological functions of the gland. Some of the actions are produced only by doses much larger than the gland could secrete during normal life.

Crude posterior pituitary extracts were first separated into two active fractions, called vasopressin and oxytocin, in 1923. The chemical structures of the active compounds were established some twenty-five years later, both being polypeptides containing eight amino-acids. In 1953, du Vigneaud and his colleagues succeeded in synthesising both oxytocin and vasopressin, a remarkable achievement which constituted the first synthesis of a polypeptide hormone. Both oxytocin and vasopressin have very similar amino-acid contents. Oxytocin contains cystine joining tyrosine, isoleucine, glutamine and asparagine as a closed ring and proline, leucine and glycine as a side chain. Vasopressin is constructed in a similar pattern except that the isoleucine and leucine are replaced by phenylalanine and arginine (Fig. 11. 12).

Destruction of the neurohypophysis, or interruption of the supraoptico-hypophysial tract (its nerve supply) which leads to atrophy of the gland, is associated in man with the disease **diabetes insipidus.** The characteristic feature of this is a marked polyuria, in which 15 litres or more of urine may be secreted in twenty-four hours ; there is a corresponding increase in thirst (polydipsia). In addition, there may be

constipation; and difficulties in parturition have been described in female patients. All these conditions may be relieved by injection of posterior pituitary extracts, though unfortunately the relief is of short duration (a few hours).

The adequacy of such replacement therapy, together with the results of surgical and other studies in experimental animals, provide evidence as to which of the pharmacological actions of pituitary extracts are to be regarded as true hormonal actions in the normal animal.

(1) **Vasopressin** is so called because it constricts blood vessels and raises the arterial pressure when injected in large doses. Its **anti-diuretic action** is exerted by much smaller doses and constitutes its most important physiological action. This action of vasopressin in reducing urine flow and increasing the concentration of the urine has

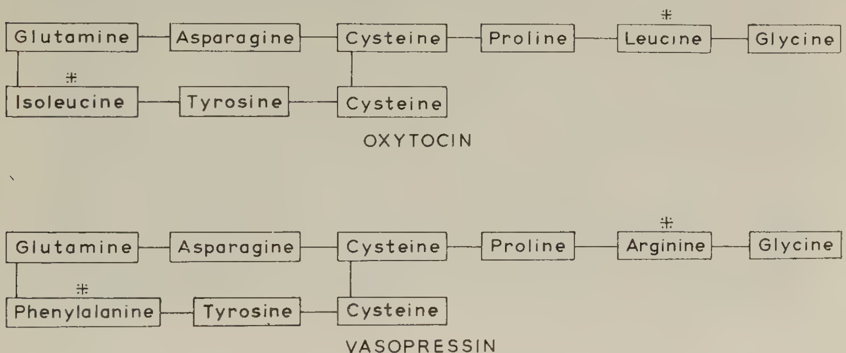


FIG. 11. 12. Composition of oxytocin and vasopressin. The nature and arrangement of the constituent amino-acids is identical in the two hormones, except that oxytocin contains isoleucine and leucine, whereas vasopressin contains phenylalanine and arginine, as is indicated by the asterisks. The two molecules of cysteine are united by the S-S bond, which closes the ring; they thus form a single molecule of cystine, and there are only eight amino-acids in the molecule.

been referred to in Chapter 9 (p. 272 and Fig. 9. 8). Experimental removal of the posterior lobe, or denervation of the gland, leads to persistent polyuria in mammals similar to that of the disease diabetes insipidus. Application of acetylcholine to, or electrical stimulation of, the supraoptico-hypophysial tract in the conscious animal inhibits water diuresis. It is presumed, therefore, that the polyuria and water diuresis are due to a lack of vasopressin, and that normally both are prevented by a steady secretion of vasopressin by the neurohypophysis. This view is largely the outcome of extensive studies of Verney and his colleagues. They have injected into the carotid artery of unanæsthetised dogs hypertonic saline and other hypertonic solutions. This produced an inhibition of water diuresis which could be accurately matched in amount and time relations by intravenous injections of vasopressin. The changes in osmolar concentration of the carotid blood passing to the head were about the same in amount (though opposite in direction)

as those involved in the systemic circulation after administration of water sufficient to produce diuresis. Moreover, injection of the same dose of hypertonic solution elsewhere into the circulation produced little or no effect since this produced a very much smaller change in concentration of the blood going to the head. It is consequently believed that the antidiuretic secretion of the gland is controlled by " osmoreceptors " probably in the neighbourhood of the supraoptic nuclei.

Substances with the properties and composition of oxytocin can be extracted from the neurohypophyses of all kinds of vertebrate that have so far been examined, except cyclostomes. But vasopressin, and the particular kind of antidiuretic action exerted by it, as just described, is peculiar to the mammals. (The vasopressin of the pig and the hippopotamus differs from that of all other kinds of mammal in that the arginine in the side-chain is replaced by lysine.)

Vasopressin is probably one of the factors concerned in the control of **intestinal peristalsis.** Electrical stimulation of the supraoptico-hypophysial tract in rabbits produces increased peristalsis and, as pointed out above, cases of diabetes insipidus may suffer from constipation.

2. **Oxytocin.** (*a*) **Uterine contraction.** The action of posterior pituitary extracts in increasing the contractions of an isolated mammalian uterus gave rise to to the term " oxytocic action " and hence " oxytocin." Lesions of the supraoptico-hypophysial tract interfere with normal birth of the young in cats and guinea-pigs. Electrical stimulation of this tract increases uterine motility in œstrous rabbits and in post-parturient rabbits and cats. Dilatation of the uterine cervix, as occurs during parturition, evokes a reflex secretion of oxytocin and an increased contraction of the uterus. Coitus, also, appears to evoke a nervous reflex release of oxytocin, and the consequent increase in uterine motility may play a part in the transport of seminal fluid up the female reproductive tract. In these experimental animals, therefore, oxytocin appears to be a true hormone. Its importance in women, however, is less certain, as discussed in Chapter 10, pp. 311–2.

(*b*) **Milk ejection.** The nervous reflex secretion of oxytocin, produced by the tactile stimulus of suckling the young, and the consequent excitation of myo-epithelial cells in the mammary gland and *ejection* of milk already present, has been discussed in Chapter 10, p. 314.

Injections of posterior pituitary extracts produce two further effects ; neither of these is of physiological importance in mammals.

(*a*) *Vasoconstriction.* When injected into anæsthetised animals, vasopressin produces a marked rise in blood pressure. The action is more prolonged than that of adrenaline, and has the curious property of rendering the blood vessels insensitive for some time (*e.g.* about half an hour) to a second injection (Fig. 11. 13). The dose of vasopressin required to raise the blood pressure is so much greater than that which arrests water diuresis, that it is unlikely that the neurohypophysis is normally related to the control of blood pressure.

(*b*) *Melanophore dilatation.* This, which leads to darkening of the

skin, occurs in amphibians (frogs and similar animals) only. It is due to contamination of the post-pituitary extract by *intermedin* from the pars intermedia, which is present in extracts of mammalian pituitary glands. It has no known action on mammals, however ; they appear to be without a tissue capable of responding to it.

The Site of Formation of the Posterior Pituitary Hormones. Microscopic examination of the neurohypophysis reveals a structure composed of nerve fibres, connective tissue, neuroglial elements and, contrary to most accounts, a rich supply of blood vessels. No cells are present

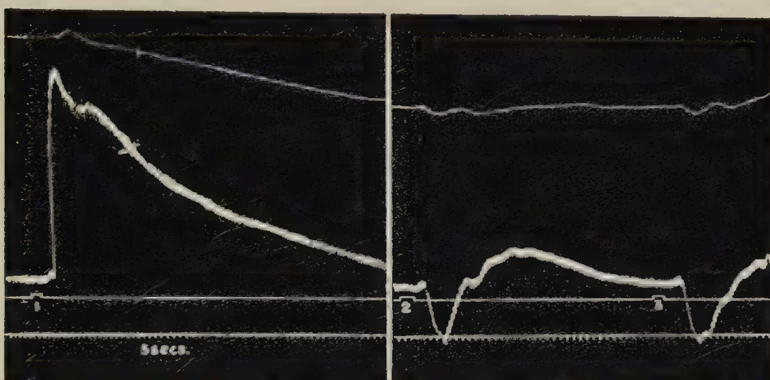


FIG. 11. 13. The Action of Extracts of the **Posterior Lobe of the Pituitary Gland** on the **Blood Pressure** of a Cat.

From above downwards: intestinal volume; arterial pressure; injection signal ; time in five second intervals. Note that the second and third injections had practically no vaso-motor action ; the fall in pressure actually produced was due to the presence of histamine as an impurity in the extract. Note also the prolonged constriction of the intestinal vessels. (Schafer and Swale Vincent.)

which have the appearance of gland cells. For this reason, it was thought at one time that the posterior pituitary hormones might be formed in the anterior lobe and transported back into the posterior lobe. Some kinds of animal, however—whales, for example—have a thick dural septum between the two lobes ; but substances with the usual activities can, nevertheless, be extracted from the neurohypophysis. There is now good evidence that the hormones are formed in the nerve cells of the supraoptic and paraventricular nuclei in the hypothalamus, by a process called " neurosecretion ": they migrate down the nerve fibres of the supraoptico-hypophysial tract and enter the blood stream in the neurohypophysis.

CHAPTER 12

NERVE

MESSAGES can be sent in the body by one of two systems, rapidly along the nerves, or slowly by hormones, "chemical messengers," in the blood. All messages from sensory receptors on the surface of the body or deep inside reach the brain along nerve fibres, and the detailed orders which the brain issues to the muscles and to most other effector organs (glands, etc.) are also sent along nerves. Hormones in the blood act more slowly and are used for more generalised and diffuse types of control.

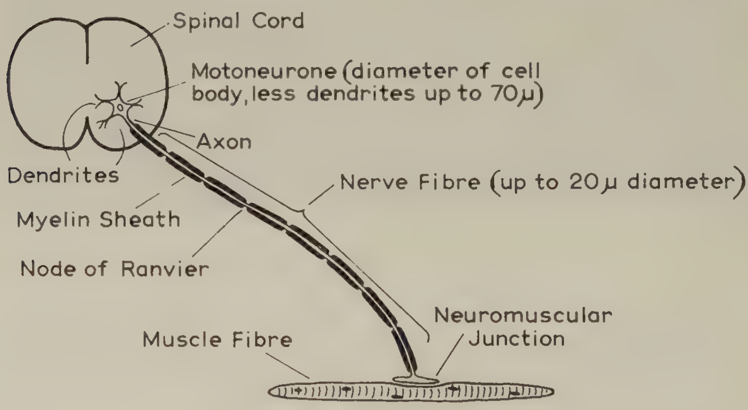


FIG. 12. 1. Diagram of a nerve fibre. A nerve fibre is a tubular process, often very long, arising from a nerve cell. The nerve cell protoplasm inside the tube is called axoplasm. In the diagram diameters are grossly exaggerated, with the true sizes roughly indicated ($1\mu = \cdot 001$ mm.).

The cell depicted is a motor nerve cell, or motoneurone, whose nerve fibre, or axon, ends on a skeletal muscle fibre. Only a single muscle fibre is shown but, in fact, a single motor nerve fibre by extensive branching inside the muscle may supply 50 or more muscle fibres. A motor nerve fibre with all the muscle fibres belonging to it is called a "motor unit."

Sensory nerve fibres, running to sense-endings in the skin and elsewhere, look precisely similar to motor fibres. Their cell bodies, however, are in the dorsal root ganglia (see Fig. 14. 1), not in the spinal cord. A peripheral nerve trunk, such as the sciatic nerve, is a bundle of large numbers of motor and sensory fibres (together with sympathetic fibres ; see Chapter 15) enclosed in connective tissue.

In vertebrates the nerve cell process is covered for most of its length by an electrically insulating fatty layer, the myelin sheath, the cell membrane being exposed only at the nodes of Ranvier, spaced about 1 mm. apart. To start with, however, the complications introduced by the myelin sheath can safely be ignored and the nerve fibre considered simply as a tube of cell membrane containing axoplasm.

The nerve fibres of invertebrates, often used for experiment because of their large size, are non-myelinated but are otherwise essentially similar in structure and mechanism to vertebrate myelinated fibres.

In the study of peripheral nerve the ultimate object is to give a complete account in terms of physics and chemistry of how signals are transmitted along individual nerve fibres (Fig. 12. 1) during normal activity in the living body. Direct investigation of single fibres in intact animals is, however, almost always impracticable. The majority of experiments have of necessity been done on isolated nerve trunks containing many nerve fibres. Fortunately there is good evidence that the mode of functioning of nerve fibres is not seriously upset when they are carefully dissected out of the body. Also the disadvantages of using a multi-fibre preparation are offset by the well-founded belief that individual nerve fibres function very independently; the sciatic nerve, for instance, contains motor fibres to very many widely separated muscles, and sensory fibres from a large area of skin. The fact that we can move a toe without any other muscles in the leg contracting, and that sensations from the toe are never confused with those from elsewhere in the leg, shows that the insulation between fibres must be excellent. These and other even stronger but more elaborate pieces of evidence give us every reason to suppose that when all the fibres in a large nerve trunk are excited together by an electric shock their individual behaviour will not be very different from usual, while the summing of all their effects makes observation much easier.

The Nerve-muscle Preparation. Many of the most important properties of excitable tissues were first discovered, and are still most conveniently demonstrated, upon a nerve-muscle preparation from a frog, usually the gastrocnemius muscle with the sciatic nerve attached. Ever since the end of the eighteenth century when Galvani discovered

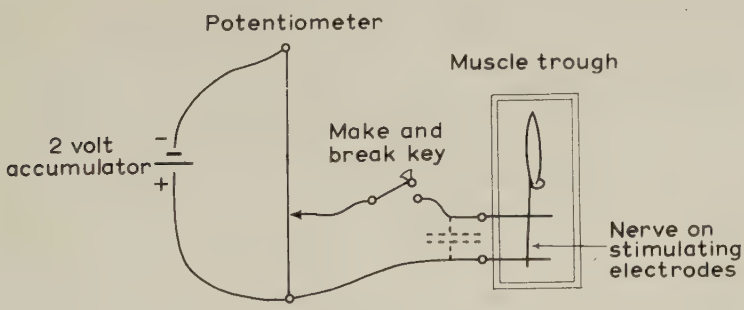


FIG. 12. 2. Apparatus for stimulation of nerve by interrupted direct current. The muscle and the nerve in the trough are covered with Ringer's fluid. Contractions of the muscle are detected by watching it.

The electrodes are of silver wire rendered non-polarisable by coating electrolytically with silver chloride. If ordinary wire is used with direct current the back electromotive force (E.M.F.) due to electrolytic polarisation of the wires causes the threshold to wander unpredictably.

To slow the rate of rise of current a condenser (shown dotted) may be connected across the electrodes. With an ordinary low resistance potentiometer a condenser of the order of 10,000 microfarads would be necessary, but smaller and more convenient values could be used if a resistance were inserted in one of the leads between the potentiometer and the condenser.

the electric current with his celebrated experiments on the excitation of the frog's leg by contact with metals, the frog's nerve-muscle preparation, stimulated electrically, has been the conventional object of neurophysiological enquiry. Many other agencies than electricity can be used to stimulate the nerve ; local heating, pinching or tapping, the application of very hypertonic solutions and of various chemical substances, and in fact almost anything which is likely eventually to injure it. But the electrical stimulus has always been preferred to these, not only because of the convenience with which its timing and its intensity and duration can be controlled, but because unless quite unnecessarily strong it appears to cause no injury to the nerve. We now know that this is no coincidence because, as we shall see later, nervous conduction is itself electrical so that the electric current is the nerve's natural stimulus.

The simplest method of exciting a nerve electrically is to pass low-voltage direct current into it, using a circuit such as that shown in Fig. 12. 2 to make and break the current and to alter its strength. With this equipment the main facts of nerve excitation are readily demonstrated.

Accommodation. In the first place it is observed that contractions can only be obtained at the moment when the stimulating current is made or broken and not when it flows continuously. This illustrates a very important general property which nerve shares with many irritable tissues: that a sudden change of stimulus is more effective than a slow or maintained alteration. We are familiar enough with this in the case of our sensations ; it is always difficult to detect a gradual change in the brightness of a light or in the loudness of a sound. With the nerve the effect of slowing down the rate of change of current when the circuit is closed can be investigated by connecting a condenser across the electrodes, as shown dotted in Figure 12. 2. A larger current is then necessary to obtain the same size of contraction as before, or, with a sufficiently large condenser, no contraction can be obtained at all. The nerve is said to " accommodate " to slowly rising currents. With peripheral nerve " slowly " means times of the order of 1/50 second ; when the switch is closed without the condenser in circuit the rate of rise of current is some hundreds of times faster than that.

Threshold. Another characteristic of nerve excitation by all kinds of stimuli is that nothing happens at all in the muscle unless the stimulus reaches a certain strength, no matter how abruptly it is applied. This is expressed by saying that the nerve has a " threshold " for the stimulus and unless the threshold is exceeded the nerve does not conduct. Such behaviour is quite different from that of many more familiar kinds of conduction ; for example, if the output from the potentiometer of Fig. 12. 2 is connected to an ordinary galvanometer, some current will flow in the galvanometer however small the voltage applied. It may be difficult to detect but it will be there all right. If, however, the nerve of a gastrocnemius-sciatic preparation is laid on the electrodes as before and the effect of making and breaking the current tried while

the voltage applied to the nerve is gradually increased, nothing whatever happens in the muscle until the potentiometer is set to give perhaps half a volt, after which the muscle contraction increases rapidly in size as the voltage is further raised. The absence of response below half a volt is genuine and not merely due to a difficulty in detecting very small contractions. No matter how sensitive the equipment no response at all can be detected below the threshold.

The threshold of a nerve is not fixed but depends, among other things, on the rate of rise of the current, and also on its duration if this is less than about 10 milliseconds (1/100 second). As the duration of the current pulse is reduced below 10 milliseconds the strength of current necessary to excite rises progressively. The experimentally determined relationship between pulse duration and threshold current is known as a strength-duration curve.

Site of Excitation. A third fundamental property is that only current leaving the nerve excites it. In the ordinary convention current flows from anode to cathode. Current will thus enter the nerve in the neighbourhood of the anode, flow along the nerve in between the electrodes, and leave under the cathode, which in Fig. 12. 2 is the electrode nearer the muscle. It is found that the nerve may be cooled, anæsthetised, or crushed either at the anode or between the electrodes without affecting the threshold when the current is switched on. But any of these operations at the cathode causes a very large increase in threshold, showing that excitation occurs under the cathode.

Break Excitation. As mentioned above the muscle also contracts when the current is broken. Similar evidence shows that in this instance excitation occurs under the anode. Anode break excitation is not a phenomenon of great significance. Many tissues, such as freshly dissected frog muscle, do not show it ; even the frog's sciatic nerve does not do so if its circulation is intact. It seems to occur particularly in deteriorating preparations in which the fibres are somewhat depolarised (see p. 363). The explanation appears to be that anodal current repolarises the fibres, but when the current stops they depolarise again rapidly and this excites them (see p. 367).

Conduction Velocity. So far we have only described how the nerve is excited and said nothing about transmission of the excitation to the muscle. Modern understanding of the nerve impulse begins with Helmholtz's demonstration in 1850 that the influence, whatever it was, that passed down the nerve from the point of stimulation to the muscle and made it contract, travelled at a definite velocity, which he measured. In the frog it is about 60 miles per hour (25 metres/second). For one form of this experiment Helmholtz invented a myograph similar to that still widely used in the classroom (Fig. 12. 3). The muscle pulled upon a lever to which was attached a long pointer writing on a revolving smoked drum. Thus a graph of contraction against time during a twitch was automatically drawn in the lampblack. The nerve was stimulated by an induction coil, it being arranged that the stimulus occurred at the same point on the tracing on each occasion. The stimulus was first applied to the nerve at a point near the muscle and a tracing

of a twitch obtained. It was then moved to the end of the nerve distant from the muscle and a twitch again recorded. The second twitch was the same shape and size as the first but the whole tracing was shifted along the drum away from the point corresponding to the time of opening of the contact breaker. Clearly the second twitch arises after a longer latency than the first and this is due to the time taken for the nervous influence set up by the shock to travel from the

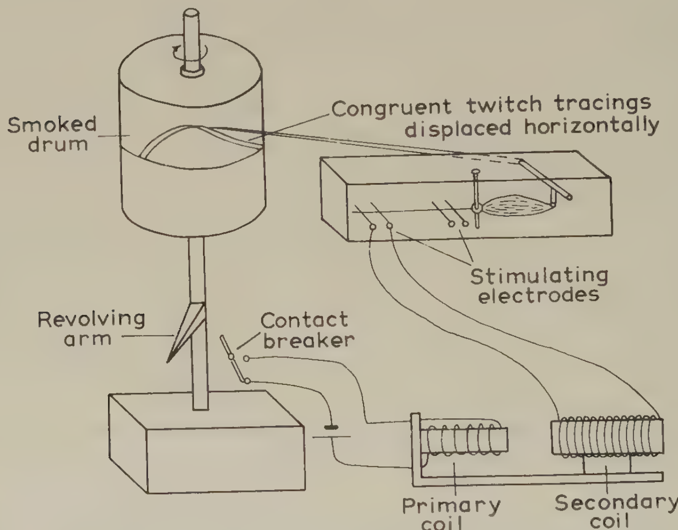


FIG. 12. 3. Measurement of the conduction velocity of frog's nerve. A sciatic-gastrocnemius preparation lies in the trough covered in Ringer's fluid. The tendon of the muscle pulls on a pivoted lever to which is attached a writing arm ; the inner end of the muscle is fixed by a pin through the knee.

Stimuli are given by an induction coil, a device similar to a motor-car ignition coil, having a primary coil of few turns and a secondary coil of many turns, which delivers a brief high voltage pulse when the current in the primary is suddenly broken by a contact breaker. The strength of shock is adjustable by altering the separation of the primary and secondary coils.

Stimuli can be applied to the nerve either through a pair of electrodes on the end of the nerve, or through another pair close to the muscle. The contact breaker of the induction coil is knocked open by an arm attached to the shaft of the myograph drum. Thus the stimulus is always delivered at the same point of rotation of the drum.

end of the nerve to the point near the muscle where the first stimulus was applied.

Helmholtz's experiment was of importance because it showed that the process of conduction in nerve was amenable to physical measurement and was not some mysterious and intangible influence passing with " the speed of thought." Helmholtz soon extended his measurements to both motor and sensory nerves in healthy human beings. For the motor fibres he stimulated ulnar nerve fibres at the shoulder and at the elbow and measured the difference in the latency of contraction of the

muscles in the ball of the thumb. For sensory fibres he measured the difference in the time a subject takes to react to electric shocks applied, say, to the foot and the loin. This method involved the unproven assumption that no factors other than nerve conduction time cause the difference in reaction time from the two sites, but the answer it gave has been confirmed by modern methods. The conduction velocities of the fastest motor and sensory fibres are roughly similar and are about 50 metres per second in human limbs. Helmholtz also discovered, on the human subject, that nerve conduction is much slowed in the cold. A 10° C. fall in temperature slows conduction by a factor of 1·7. Human limbs are often surprisingly cold in winter and even in summer seldom reach " body temperature " (37° C. or 98·4° F.).

The Resting Potential. At about the same time that Helmholtz was measuring conduction velocities, du Bois Reymond and others were investigating the electrical phenomena of nerve and muscle which led eventually to the electrical theory of nerve conduction. It was found that when a nerve or muscle was cut across, the cut end showed a negative potential difference of a few hundredths of a volt relative to the intact tissue. This is because the interior of nerve and muscle fibres is negative to the exterior with a potential difference across the cell membrane ; normally the cell membrane surrounds the whole fibre,

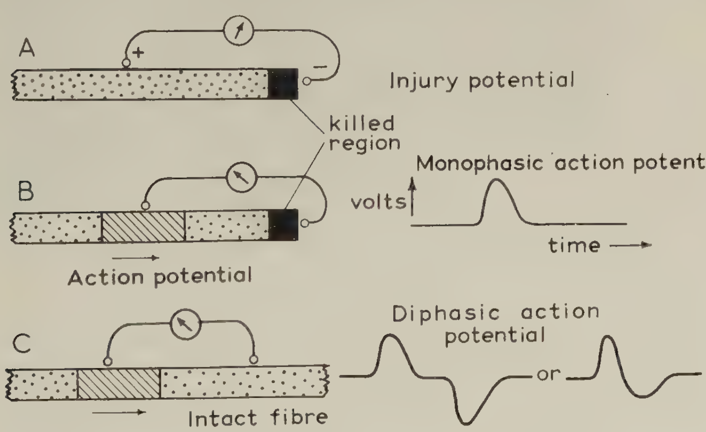


FIG. 12. 4. Mode of recording injury and action potentials from a nerve fibre.

A. When the fibre is cut across the killed end shows a negative potential difference with respect to the intact part.

B. As an action potential (itself a region of negativity) passes under the electrode on intact nerve it causes a brief reduction in the potential difference. The appearance of the action potential recorded in this way on a cathode ray tube oscilloscope is shown to the right. The action potential fades out in the killed region and does not reach the end of the fibre. Hence the action potential is only detected as it passes the electrode on intact nerve, and the record is " monophasic."

C. With electrodes on an intact fibre a potential change, first in one direction and then in the other, is recorded as the action potential passes under each electrode in turn. This gives a " diphasic " record. Often the two phases run into each other giving the appearance shown on the extreme right.

so that if two electrodes are placed on the surface at different points no potential difference is found. But cutting the fibre allows contact to be made with the interior and so we get the "injury potential" (Fig. 12. 4). The magnitude of the injury potential recorded in this way is less than the true resting potential because, at the cut end, fluid and dead tissue, inevitably present, provide a current path between the inside and outside of the fibre which short-circuits the resting potential and locally depolarises the fibre. The electrode on the cut end therefore

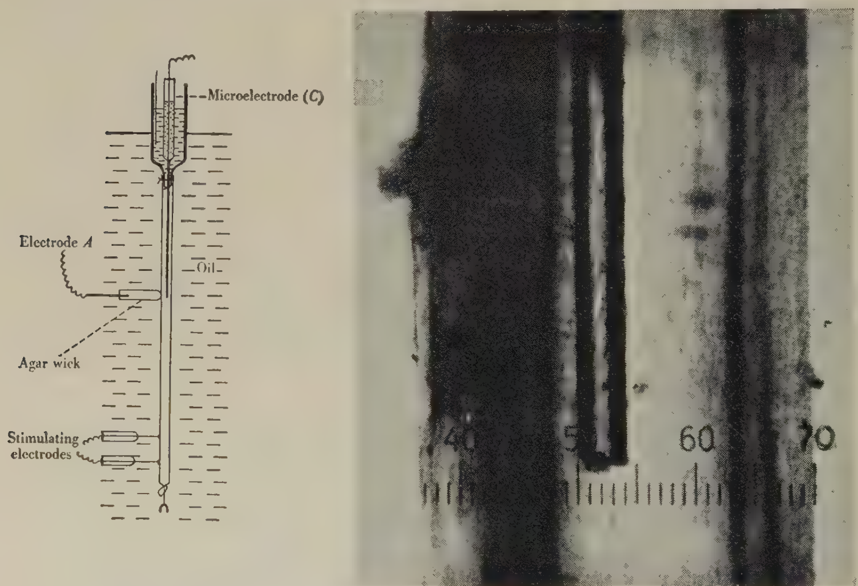


FIG. 12. 5. The arrangements for recording resting and action potentials from a giant nerve fibre of a squid.

A cannula is inserted into the cut end of the axon, which hangs freely in oil with a small weight tied to the bottom end. A glass capillary microelectrode (C) is inserted through the cannula as far into the interior of the axon as is desired, usually several centimetres. Contact with the outside of the axon is made by means of a wick soaked in sea water (electrode A).

The photograph shows a microelectrode inside a living squid axon. The axon is seen as a clearer band surrounded by undissected connective tissue and smaller fibres. Each division of the graticule equals 33 μ; the axon is thus 500 μ ($\frac{1}{2}$ mm.) in diameter and the electrode 130 μ. (Hodgkin and Huxley, *Nature*, vol. 144, p. 710, 1939.)

makes contact with the axoplasm at a point where the potential difference is smaller than the true resting potential.

The true resting potential can be measured by thrusting a fine glass capillary electrode along the inside of an axon until the tip is well clear of the depolarised region. This was first achieved in 1939 using giant axons from the squid. The method used is illustrated in Fig. 12. 5 and a record of potential made in this way is shown in Fig. 12. 6. The resting potential of isolated squid nerve is about 50 millivolts (·05 volt).

For fresh vertebrate nerve and muscle fibres the value (obtained by analogous methods) is about 90 millivolts.

The Action Potential. du Bois Reymond discovered that when a nerve is stimulated, the resting potential, recorded at the other end of the nerve, diminishes. At the time only slow galvanometers were available, but when rapid recording instruments were developed it was found that the reduction was due to a brief wave of negativity which travelled along the nerve from the stimulating electrodes. As this

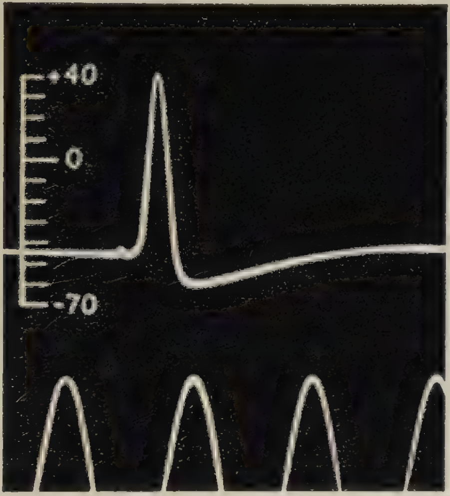


FIG. 12. 6. Resting and action potentials from a single nerve fibre of the squid *Loligo forbesi*, recorded with an internal microelectrode by the method illustrated in the previous figure. The potential difference between the inside and outside electrodes has been amplified by a direct-coupled valve amplifier and displayed on a cathode ray tube oscilloscope.

The vertical scale indicates the potential of the internal electrode in millivolts (mV), the sea water outside being taken as zero potential. Time markers at 2 millisecond intervals.

The record reads from left to right. The potential to start with is the resting potential of − 45 mV. The first small deflection of 2–3 mV is an artefact caused by the stimulus. It is followed by the action potential which overshoots zero and rises to + 40 mV. (Hodgkin and Huxley, *Journal of Physiology*, vol. 104, p. 176, 1945.)

wave passed under the electrode which was on intact nerve, it caused a temporary reduction in the potential difference between that electrode and the one on the cut end (Fig. 12. 4). This electrical wave is called the action potential or, because of its shape, the spike potential. It used to be thought that during the action potential the resting potential was abolished, so that an electrode over the active region and one on the cut end would ideally record zero potential difference. Hence the action potential was spoken of as a wave of depolarisation. In 1939 it was discovered that, when true potential differences are recorded by an electrode inside the nerve fibre, the action potential is larger than

the resting potential (Fig. 12. 6). It not only cancels the resting potential but overshoots by some 30–40 mV. so that the inside becomes actually positive to the surface of a region at rest. The implications of this revolutionary observation will be considered later.

The Class of Disturbance to which the Nerve Impulse Belongs. The All-or-none Law. A very important question which has attracted much interest from the earliest days is the relationship between the nervous impulse (defined as whatever process it is that passes down the nerve and makes the muscle contract) and the action potential. Are they the same? In spite of much experimentation no one has ever been able convincingly to separate the two. It has been shown that they have the same threshold, propagate at the same velocity, are blocked at the same time if the nerve is frozen or compressed, and so on. Evidence of this kind shows that the nerve impulse and the action potential are so closely related that they cannot be dissociated; it does not, unfortunately, prove they are the same, although many people made this suggestion. It permits, and perhaps encourages, this theory, but would be equally consistent with the view that the essential process is some chemical reaction which is always accompanied by electrical changes; the action potential would then be, as it were, merely the noise of the engine. Before describing Hodgkin's experiments which refuted this last possibility and proved that nervous conduction is electrical it will be convenient to bring forward various other facts which throw light on the kind of process that nervous conduction is, and which show how the particular hypothesis of conduction that he tested came to be proposed.

The action potential is, as we have seen, a travelling wave of electric change; in frog's nerve it moves at about 2,500 cm./sec. and lasts about 1/1,000 of a second, so that the active region is confined at any instant to about $\frac{2{,}500}{1{,}000} = 2{\cdot}5$ cm. length of nerve. Hence in a nerve as long as the sciatic of the frog, stimulated at the end, all activity has ceased at the point of stimulation well before the action potential arrives at the muscle. The action potential is, thus, quite a localised affair and nervous conduction must be wholly unlike what happens in an ordinary telegraph wire; the common analogy drawn between nerves and telegraph wires is only very superficial.

As this region of electrical disturbance moves down a long nerve in uniform surroundings it neither decreases in size, nor travels more slowly. This suggests that the nerve impulse obtains the energy it needs for propagation locally as it goes along and not, like a rifle bullet, from the stimulus that starts it. This fits in with older observations that the stimulus does not have to be larger when it is applied to the nerve further away from the muscle, and that the impulse cannot be made to go faster by using a larger stimulus than is necessary just to excite.

Not only does the energy for propagation appear to be acquired locally, but local conditions also determine its rate of release. If a section of nerve is cooled the impulse goes more slowly in that length,

but speeds up again to its original rate when it reaches warm nerve again. Thus the conduction velocity in each stretch is determined only by the conditions in that stretch and not at all by the previous history of the impulse.

Such behaviour would be most simply explained if the impulse was self-propagated by the active region exciting the next section ahead, like a flame passing along a train of gunpowder. The characteristics of the response at any point would depend only on local conditions and not on the response of the next door sections that excited it. This theory is clearly a good one to explain the above observations on conduction velocity but it carries the implication that the size as well as the velocity of the impulse ought to be independent of the size of the stimulus. If release of energy for propagation is a purely local affair, one of the things determined locally should be the size of the impulse. At any point either you should get a full-sized impulse or none at all ; it should not be possible to get a half-sized impulse by carefully grading the size of the stimulus. Behaviour of this kind is referred to as " all-or-none," and tissues which display it are said to obey the " all-or-none law." All-or-none behaviour was first observed in the contraction of the heart by Bowditch in 1872.

On the face of it, nerve does not obey the all-or-none law. As we have said earlier, if the nerve of a nerve-muscle preparation is stimulated with stimuli of increasing strength, first threshold is reached, but above threshold the contractions of the muscle for some time increase in size as the stimulus increases ; so the preparation as a whole clearly does not behave in an all-or-none manner. An obvious explanation is that each individual motor nerve fibre does obey the all-or-none law, but because the nerve to gastrocnemius contains many fibres with different thresholds they are excited one after another and so the relationship is concealed. This explanation was shown to be correct by Keith Lucas in 1909. He simplified the problem by using a very small muscle, the cutaneous dorsi muscle of the frog, which has only 8 or 9 motor nerve fibres. When the motor nerve is excited with steadily increasing strengths of shock the size of the contraction resulting does not increase smoothly but in a series of abrupt steps (Fig. 12. 7). The number of steps is never more than the number of motor nerve fibres to the muscle. Each step clearly represents the excitation of one of the motor fibres, and the fact that each fibre has a sharp threshold, and that there is no further increase in the size of contraction until the next step is interpreted to mean that, in each fibre, the size of the impulse does not increase with increasing shock size, once the threshold is passed. This experiment, it should be noted, establishes the all-or-none law for nerve, not muscle. The muscle fibres to which the nerve fibre is connected only come in as indicators of the arrival of a nerve impulse.

Nowadays the all-or-none law is readily observed on the single giant nerve fibres of invertebrates. With the arrangement shown in Fig. 12. 5 it is found that, as the stimulus is increased from zero, at first the trace on the cathode ray tube remains quite flat, except for a

small stimulus artefact, until suddenly the picture shown in Fig. 12. 6 appears. Further increase in the stimulus causes no increase in the height of the action potential. No amount of adjustment of the stimulus strength ever results in a half-sized action potential. Thus the size of action potential is independent of the stimulus strength once this has reached threshold.

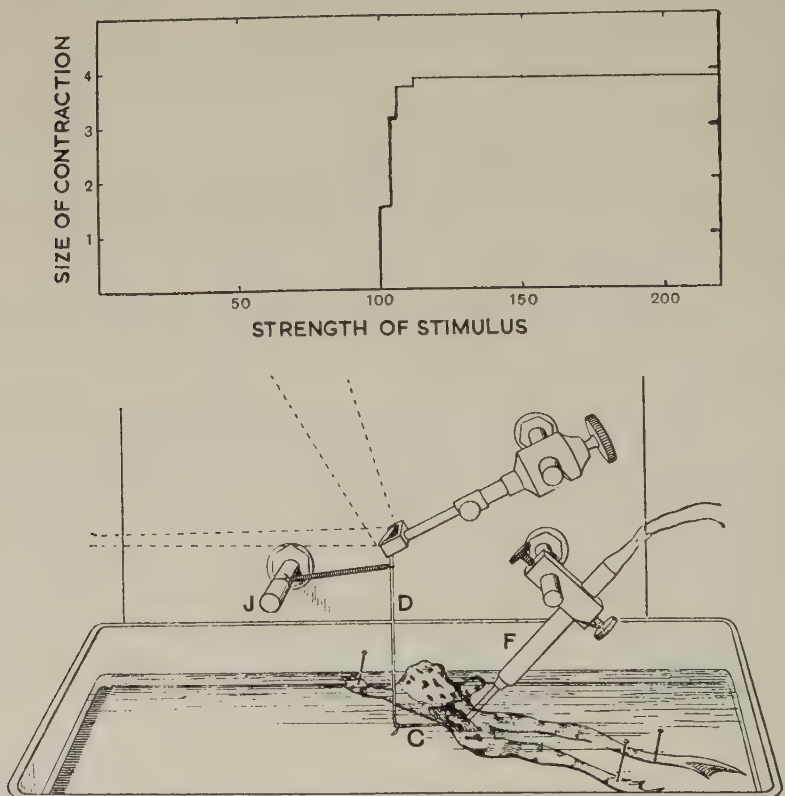


FIG. 12. 7. Keith Lucas's apparatus for demonstrating the all-or-none law, using the dorso-cutaneous muscle of the frog.

The piece of skin into which the muscle, C, is inserted is cut free and attached to the lever, D, which carries a mirror. The muscle is lightly stretched by a spring between the lever and the support J. F, stimulating electrodes applied to the motor nerve. Contractions are recorded photographically by means of a beam of light reflected from the mirror.

Above is shown the relationship observed between strength of stimulus (threshold taken as 100) and the size of contraction. The number of steps was never more than the number of motor nerve fibres. (*Journal of Physiology*, vol. 38, p. 113, 1909).

The all-or-none behaviour of the nerve impulse is important not only because of the light it throws on the nature of nervous transmission, but because it imposes the signalling code for the whole nervous system. Information and instructions can only be sent along a nerve fibre by alter-

ing the frequency of nerve impulses and not by altering their size. This system of frequency modulation rather than amplitude modulation, to borrow radio terminology, has very great advantages for long distance signalling—as radio engineers have discovered. It makes nerve signalling independent of local conditions ; all that counts is the number of impulses that get through, and it does not matter if cold or pressure or shortage of oxygen slows them up or reduces their size, as long as conduction does not fail altogether. Likewise it is responsible for the independent functioning of the many fibres bundled together in nerve trunks, for although the electric current generated by an impulse in one fibre will reach its neighbours, it is far too small to excite impulses in them (let alone to extinguish impulses already there) and unless it does this it will not interfere with their signalling at all.

The Local Circuit Theory of Transmission. The demonstration of its all-or-none behaviour removed the last difficulties in regarding the

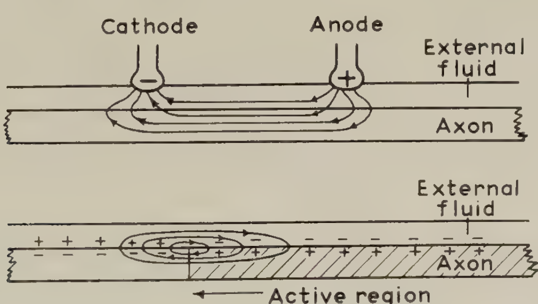


FIG. 12. 8. The local-circuit hypothesis of nervous transmission.

The upper diagram shows a single nerve fibre, covered with a layer of fluid, with two stimulating electrodes. Current enters the fibre near the anode and leaves near the cathode, in the manner shown by the lines of current flow.

In the lower diagram an action potential, a region of reversed membrane potential, is travelling from right to left. The surface of the active region is negative with respect to the surface of the resting nerve ahead. Current therefore leaves the resting nerve and flows in the outside fluid down the potential gradient to the active region. The local circuits are completed via the axoplasm, where the same argument applies in reverse.

Thus in front of the active region current leaves the nerve fibre, just as it does under a stimulating cathode. These currents will tend to excite the nerve in front of the active region. The hypothesis is that they do in fact do so, and that this is how the action potential propagates.

nerve impulse as a self-propagating disturbance. We have now to consider by what mechanism the active region excites the region ahead. Many possibilities are ruled out by the observation that nerve fibres (whether sensory or motor) transmit equally well in either direction. If a length of nerve is excited in the middle the action potential spreads in both directions with equal velocity. Whatever it is that excites locally during propagation must therefore pass with equal ease up or down the fibre. Either diffusion of a chemical substance or the flow of electric current might serve ; but diffusion looks like being too slow a process

to account for conduction at 60 miles per hour, while there are obvious reasons for preferring an electrical hypothesis ; electric currents very easily excite ; the action potential produces an electric current ; moreover the polarity is correct, for it is the cathode which excites and likewise the active region is negative.

The hypothesis is, then, that propagation occurs because the active region causes local currents to flow in the inactive region ahead (Fig. 12. 8). These currents are in the same sense as the currents that flow when a cathodal stimulus is applied from an external electrode and excite for the same reason. There are no observations which are inconsistent with the local circuit theory and since the end of the nine-

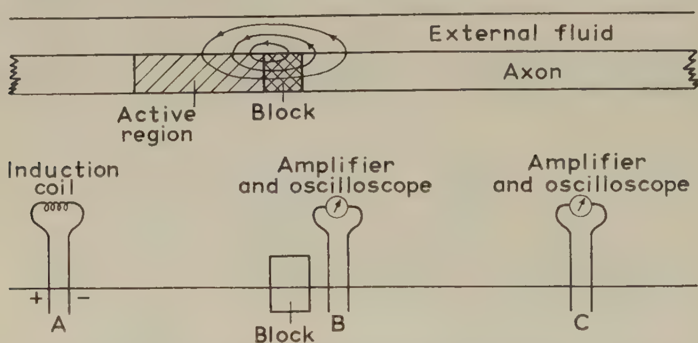


FIG. 12. 9. Evidence for electrical transmission in nerve.

A frog's sciatic nerve is blocked by freezing a short length. When the nerve is excited by shock from an induction coil applied through the electrodes at A, no action potential gets through to the recording electrodes at C, but immediately after an action potential from A arrives at the blocked region electrodes at B record small local currents spreading through the block. The upper diagram shows, for one fibre, how current spreads through the blocked region, just as it spreads in front of a normal impulse (compare with Fig. 12. 8).

The electrodes at B can also be connected to an induction coil and used to measure the threshold of the nerve, the appearance of an action potential at C indicating that threshold has been reached. Immediately after an action potential arrives at the other side of the block a transient fall in threshold can be detected at B. The fall in threshold occurs at the same time and has the same spatial distribution along the nerve, as the local currents set up by the blocked impulse. (After Hodgkin, *Journal of Physiology*, vol. 90, p. 183 and p. 211, 1937.)

teenth century it has been the most widely accepted hypothesis of propagation, but until 1937 no one knew whether it was true or not. Certainly the local currents flow in the right way to excite, but the unanswered and critical question was: Are they in fact large enough to do so ?

Hodgkin solved this question in the following way. In a frog's sciatic nerve he blocked conduction by freezing a short length (Fig. 12. 9). Immediately after an action potential had arrived at the block, small currents of the kind expected from the local circuit theory could be detected spreading several millimetres beyond the block. Test stimuli showed that at the same time the threshold of the nerve in

this region fell sharply. On occasion, the threshold was reduced by as much as 90 per cent. If the local currents in front of a blocked impulse can lower the threshold by 90 per cent. those from a normal impulse would certainly lower it by 100 per cent., *i.e.* they would excite. Hence the action potential does constitute a large enough electrical stimulus to excite the nerve ahead of it. It is now known, in fact, that the normal action potential is five to ten times larger than it has to be in order to excite. Electrical propagation has a large safety factor.

Conduction Velocity. The velocity of propagation depends on the strength of current flowing in the local circuits ahead of the active region. As described above (p. 349) any current must flow for a certain time in order to excite ; the larger the current the shorter the time for which it has to flow. Hence the larger the local currents, the sooner the section of nerve ahead of the active region is excited and the higher the conduction velocity. The strength of the local currents depends, amongst other things, on the electrical resistance of the outgoing path (*i.e.* the longitudinal resistance of the axon) and the resistance of the return path through the external fluid. The longitudinal resistance of the axon decreases as the cross sectional area of the axoplasm increases. Thus large nerve fibres conduct faster than small. The squid escapes from its enemies by contracting its mantle, which drives it rapidly backwards through the water. To do this as fast as possible it has developed its giant nerve fibre of 500 μ diameter, which conducts at 20 metres/sec.

In the body nerve fibres are effectively surrounded by a large volume of fluid, so that there is little scope for increasing conduction velocity by lowering the resistance of the return path. In isolated fibres, however, the conduction velocity can be altered by changing the resistance of the return path, an experiment which provides powerful evidence for the electrical theory of propagation. Hodgkin laid a giant axon in air on a grid of platinum strips ; electrodes were also provided for stimulating at one end and recording at the other, so that the conduction velocity could be measured. The ends of the platinum strips could be plunged into a trough of mercury, in order to connect them all together electrically. This had the effect of suddenly reducing the longitudinal resistance outside the nerve and, at the same moment, the velocity of the action potential was seen to increase. The fact that conduction velocity can be altered instantaneously by making electrical connections right away from the nerve would be almost impossible to explain unless conduction was itself an electrical process.

The Ionic Basis of Nerve Action

The foregoing experiments establish that the action potential propagates by local electric circuits, but although they tell us how it propagates they reveal nothing about the nature of the potential change itself. By what mechanism does the potential difference across the nerve membrane suddenly reverse and almost equally rapidly return to its resting state again, as it is seen to do in Fig. 12. 6 ? The answer

is that the reversals of potential are due to movements of charged sodium and potassium ions, across the nerve membrane. As already described in Chapter 8, the resting potential, from which the action potential takes off, is itself a consequence of the unequal distribution of sodium and potassium ions across the nerve membrane. Before describing how the potential difference across the membrane is reversed by ionic currents during the action potential, it is first necessary to look into the nature of the resting potential in greater detail.

The Resting Potential. In sea water and in the blood and extracellular fluid of animals the chief ionic constituents are sodium and chloride ions. There is some potassium, but potassium ions only represent about 2·5 per cent. of the total cations. Inside living cells, with few exceptions, the situation is reversed ; there is a lot of potassium

TABLE 12. 1

Approximate concentrations of potassium, sodium and chloride in nerve and muscle fibres and in the fluid bathing them (artificial sea water or Ringer's fluid). 1 m.mole/l = 10^{-3} gram-molecules per litre.

Tissue	Potassium		Sodium		Chloride	
	Inside m.mole/l.	Outside m.mole/l.	Inside m.mole/l.	Outside m.mole/l.	Inside m.mole/l.	Outside m.mole/l.
Squid axon	400	10	50	460	110	540
Frog nerve and muscle	120	2·5	15	120	3	120
Mammalian muscle .	140	4	10	150	—	140

(roughly as much as there is sodium outside) but little sodium and chloride (Table 12. 1). Although there is this small amount of chloride, by far the greater part of the internal anions consists of organic molecules (in the case of squid nerve mainly amino acids and other small organic acids) carrying net negative charges. Thus, outside cells, we have chiefly sodium ions (Na^+) and chloride ions (Cl^-) and inside, potassium ions (K^+) and organic anions. In solutions of such strength, sodium chloride is completely ionised and does not exist as a compound. Similarly there is good reason to believe that all, or very nearly all, the internal potassium also exists as free ions. It is misleading and, in a sense, incorrect to speak of cells as containing " potassium proteinate," as is sometimes done.

Leaving to one side the question of how these striking differences in the composition of the cells and the fluid bathing them arise in the first place, let us accept them and consider how they are maintained. One very simple way of keeping the K^+ in and the Na^+ and Cl^- out would be to have an impermeable cell membrane. But such cells as have been examined prove to be permeable to K^+ and Cl^-, although they are much less permeable to Na^+ and the organic ions. It might be thought that if the cell membrane were permeable to K^+ and Cl^- the

only way to keep the K^+ concentration high and the Cl^- low inside the cells would be to pump the K^+ back as it leaks out and pump the Cl^- out as it comes in. But this is not the only way, for the tendency of the K^+ to diffuse out can be offset by an electrical potential difference across the membrane. As potassium ions are positively charged, the inside of the cells must be negative relative to the outside in order to attract any potassium ions that try to leave. Since chloride carries an opposite (negative) charge the same potential difference will serve to keep out the chloride ions too. In fact in all cells on which measurements have been made (mainly nerve and muscle) a resting potential of roughly the expected size has been found. Thus K^+ and Cl^- are (to a first approximation) in equilibrium across the membrane, the tendency of the ions to diffuse down the concentration gradients across the membrane being balanced by the potential gradient.

The value of the "equilibrium potential" which just balances the tendency for a univalent ion of internal concentration C_i to diffuse out into an external fluid containing a concentration C_o is given by the expression $E = \frac{RT}{F} \log_e \frac{C_i}{C_o}$ where T is the absolute temperature, R is the gas constant (as in the gas equation PV = RT), and F is Faraday's constant (96,500 coulombs, the amount of electricity needed to deposit 1 gram equivalent of a substance at an electrode). Inserting numerical values of R, T and F and converting to logarithms to base 10 we get :

$$E \text{ (in millivolts)} = 58 \log_{10} \frac{C_i}{C_o} \text{ at } 20° \text{ C.}$$

In fresh squid nerve and in vertebrate nerve and muscle the ratio $\frac{C_i}{C_o}$ for potassium is about 35, giving a potassium equilibrium potential of 58 log 35 = 90 mV., which is not far from the observed values of resting potential.

An important point that must be grasped in order to understand the ionic mechanism of nerve is that relatively very few ions indeed have to move in order to produce the changes of potential involved. The potential change caused by a certain movement of charge is proportional to the capacity of the structure into which it flows (1 coulomb flowing into a capacity of 1 farad changes the potential by 1 volt). The nerve membrane has a capacity of about 1 microfarad per square centimetre and the resting potential is roughly 100 mV. = 1/10 volt. To change the potential difference of 1 sq. cm. of membrane by 1/10 volt therefore needs 1/10 microcoulomb = 10^{-7} coulombs of electricity. Now 1 gram molecule (mole) of a univalent ion carries a charge of 96,500 coulombs, say 10^5 coulombs, so that 10^{-7} coulombs corresponds to only 10^{-12} mole of a univalent ion. In a 500 μ squid axon the volume of axoplasm covered by 1 sq. cm. of membrane is approximately 1/100 ml. Squid axoplasm contains some 0·4 mole of potassium ions per litre so that amount of potassium in 1/100 ml. is $0{\cdot}4 \times 10^{-5}$

mole. Thus the axon contains $\frac{0 \cdot 4 \times 10^{-5}}{10^{-12}}$ = 4 million times more potassium ions than would have to pass across the membrane to alter the potential difference across it by 100 mV.

We are now in a position to understand by means of a hypothetical model just how the resting potential arises (Fig. 12. 10). Let us in imagination make up something to represent extracellular fluid by dissolving some sodium chloride and a much smaller amount of potassium chloride in water. The inside of the cell will be represented by a solution of the potassium salt of some amino acid (for the sake of definiteness say aspartic acid), together with a little sodium chloride. These solutions

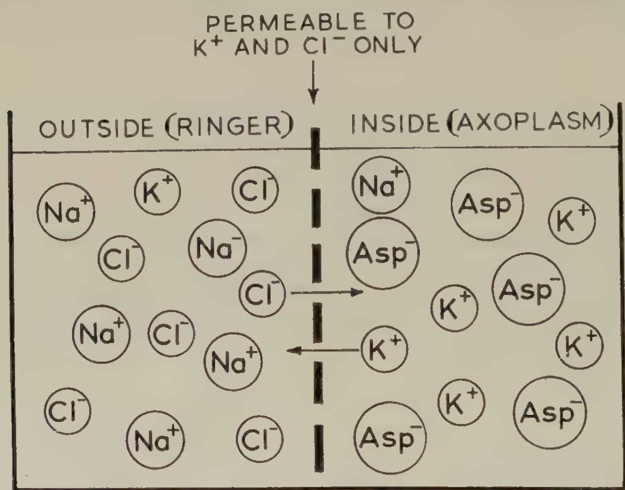


FIG. 12. 10. Model illustrating the origin of the resting potential. A hypothetical partition permeable to K^+ and Cl^- but not to Na^+ or to organic ions, represents the nerve membrane. The solution in the compartment labelled " outside " is a simplified Ringer's fluid of NaCl with a small proportion of KCl. Axoplasm " inside " is represented by a solution of potassium aspartate with some NaCl, isotonic with Ringer.

are made up from laboratory chemicals in separate beakers. We now pour the two solutions into a chamber divided into two parts, which we will call inside and outside, by a membrane permeable to K^+ and Cl^-, but not to Na^+ or aspartate. Because of the concentration differences K^+ at once begins to pass from the inside to the outside and Cl^- from outside to inside. Owing to their large charge, however, few ions have to pass before a potential difference is set up large enough to oppose these movements. Thus a membrane potential rapidly builds up, but the movement of ions necessary is so small that if samples of the two fluids were withdrawn and analysed no change in composition could possibly be detected. The excess of anions over cations inside necessary to give the inside a negative potential relative to the outside would be, as we have seen, of the order of a millionth part of the total.

All cells that have been investigated so far have resting potentials in which the interior of the cells is negative to the surroundings. It is important to remember that this fact could not have been foreseen from a knowledge of the ionic compositions of the cell contents and the external fluid. If instead of being permeable to K^+ and Cl^- and impermeable to Na^+ the reverse had been true, there could have been an equally good equilibrium but the inside would have been positive. This can easily be understood with the above model. When the solutions are poured into the compartments, Na^+ now being the only ion that can move, some sodium ions immediately move inwards. As they carry a positive charge the inside becomes positive and the outside becomes negative. Soon the potential difference becomes sufficient to prevent any further movement of Na^+ down the concentration gradient and equilibrium is established with the inside positive.

This situation is not of merely hypothetical interest. As we shall see later, at the height of the action potential the nerve membrane does become predominantly permeable to Na^+ and the potential difference across it is close to the equilibrium potential for Na^+, with the inside positive. Secondly even at rest the membrane is slightly permeable to Na^+. This complicates matters. The simplest way of looking at it is to regard the resting membrane as consisting of a large area of membrane permeable to K^+ and Cl^- and a small area permeable to Na^+. Each area tries to reach its own equilibrium potential but fails. The resultant potential is near the equilibrium potential for K^+ and Cl^-, but shifted some millivolts towards the Na^+ potential, *i.e.* the inside is less negative. Such a system cannot be in equilibrium. Since the membrane potential is less than the K^+ equilibrium potential, K^+ will diffuse out ; and similarly Na^+ enters as the potential is very far from the Na^+ equilibrium potential. Left to itself such a system would slowly run down until the ionic concentration differences levelled out. That it does not do so in living cells is due to an active pump, driven by metabolic energy, which pumps Na^+ out and K^+ in. Thus cells maintain a steady state, but they do not have a simple membrane equilibrium, because energy has to be supplied to maintain their ionic composition.

In general, the potential across the membrane at any time depends on the relative permeabilities to Na^+, K^+ and Cl^-. Other ions to which the membrane is permeable are present in much lower concentrations and make little impression. Normally the membrane of nerves and muscles is moderately permeable to K^+ and Cl^- and almost impermeable to Na^+. So the resting potential is close to the K^+ and Cl^- potential. At the height of the action potential the membrane becomes extremely permeable to Na^+ and the potential difference is the Na^+ potential. Other conditions are also found, as we shall see subsequently ; thus the muscle end-plate under the action of acetyl choline becomes very permeable to both Na^+ and K^+ and the membrane potential therefore nearly vanishes.

The idea that the resting potential arises because of the K^+ concentration difference was long ago suggested by the fact that raising the

external K^+ concentration reduces the injury potential. It was also known that muscle fibres were permeable to K^+ but not to Na^+ ; they take up K^+ if the external concentration is raised but not Na^+. Thus if a muscle is placed in hypertonic sodium chloride it shrinks permanently, but if the bathing solution is made hypertonic by dissolving solid potassium chloride there is temporary shrinkage but the fibres soon regain their original volume and are found to have taken up sufficient extra K^+ and Cl^- to re-establish osmotic equilibrium. But quantitative evidence was lacking, partly because the absolute value of the resting potential was uncertain, and partly because in conventional tissues such as frog nerve and muscle it is difficult to get reliable estimates of the Na^+ and Cl^- concentrations inside the fibres ; the concentrations inside are rather small, but very large in the extracellular fluid, so that a small error in estimating the amount of the latter produces very large errors in the internal concentrations.

Since 1945, work on single fibres, principally the squid giant axon, has given the necessary data. The permeabilities of the membrane to Na^+ and K^+ have been measured by observing the rate of uptake and of loss of radioactive Na^+ and K^+ by the axon. Furthermore the axoplasm of squid nerve can be squeezed out from the cut end like toothpaste from a tube, uncontaminated by extracellular fluid, so that accurate values for internal Cl^-, Na^+ and K^+ have been obtained. The value for the resting potential calculated from the permeabilities of the membrane to Na^+ and K^+ and from the concentrations of these ions inside and outside the axon, given in Table 12. 1, agrees closely with the actual value, measured with an internal electrode. (The measured internal Cl^- concentration, however, is greater than it should be if the distribution of Cl^- is determined only by the resting potential. The reason for this is not yet understood.)

Any possible doubt that the resting potential is determined by the concentrations of the ions and the passive permeability properties of the membrane was finally removed when it proved possible to squeeze all the axoplasm out of a squid axon and then to re-inflate it and perfuse it with solutions of known composition. With sea water outside and an isotonic solution of a potassium salt inside, the resting potential has normal values (50–70 mV.) and the axon conducts action potentials of normal size for many hours. With isotonic potassium salt outside and isotonic sodium chloride inside the resting potential reverses, the inside of the nerve becoming 50–60 mV. positive to the external fluid.

The Sodium-Potassium Exchange Pump. The mechanism which pumps sodium out of cells and potassium in, was first discovered in red cells. In blood that is stored in a refrigerator the red cells gradually lose K^+ and gain Na^+. When they are incubated with some glucose the Na^+ content falls again and the K^+ rises. As these movements are against the concentration gradients and must involve doing osmotic work it is clear that an active pump using metabolic energy is involved. A similar mechanism has been shown with great clarity in giant single

nerve fibres from the squid. Hodgkin and Keynes pushed a fine glass capillary into a squid axon in the usual way but, instead of recording potential differences, it was used to inject a very small quantity of radioactive Na^+. Within a few seconds radioactivity began to appear in the fluid bathing the fibre. Most of this loss of radioactivity was due to active extrusion of Na^+, for the rate of loss dropped steeply if the

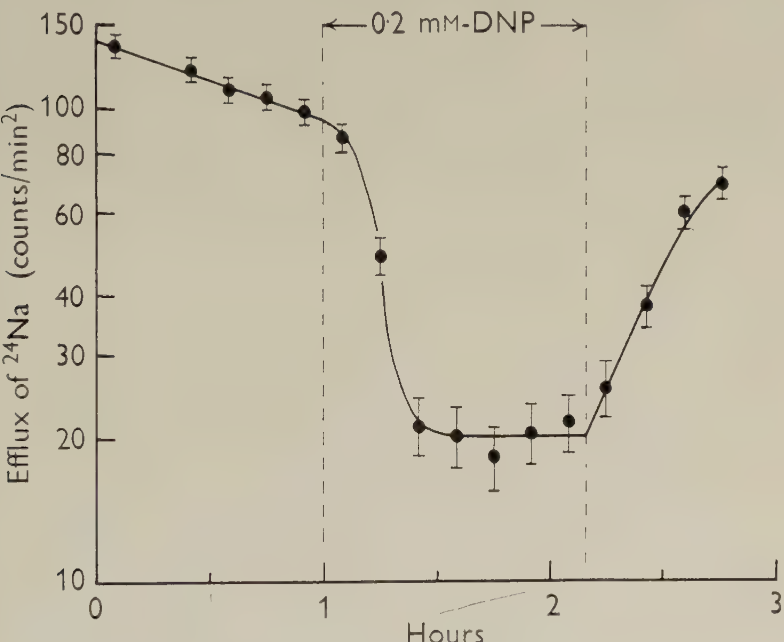


Fig. 12. 11. Extrusion of radioactive sodium by a squid giant axon.

An internal microcapillary similar to that shown in Fig. 12. 5 was used to inject a small quantity of solution containing ^{24}Na into the axoplasm. The sea water bathing the outside of the fibre was changed at 10 min. intervals and the amount of radioactivity in successive samples, due to extruded ^{24}Na, measured in a Geiger counter. Quantity of radioactivity is expressed in counts/min. and the units of rate of efflux are, therefore, (counts/min)/min.

The addition of 0·0002 gram-molecules per litre (0·2 mM) dinitrophenol (DNP) to the sea water outside the fibre reduced the efflux by a factor of 5. The pump recovered when the DNP was washed away. (Hodgkin and Keynes, *Journal of Physiology*, vol. 131, p. 592, 1956.)

pump was poisoned with dinitrophenol (Fig. 12. 11) or by cyanide or azide.

Energy for the pump comes from hydrolysing the phosphate bonds of "energy-rich" phosphate compounds such as adenosinetriphosphate (ATP). The above poisons stop the production of energy-rich phosphate compounds in the mitochondria. The pump can be restarted again by an injection of ATP down the internal microcapillary. The pump is also poisoned by the cardiac glycosides, the digitalis group of drugs,

but in this instance it is not the formation but the utilisation of energy-rich phosphate compounds that is interfered with. The cardiac glycosides are known to inhibit some adenosine triphosphatases.

At one time it was thought that the pump need only extrude Na^+. If we imagine a cell initially containing a lot of Na^+ and Cl^- and a little K^+, similar, that is, in composition to the outside fluid, and proceed to pump out the Na^+, then removing the Na^+ will leave the inside negative, so that K^+ will be drawn in and Cl^- pushed out. Thus by merely pumping Na^+ a situation not unlike that found could certainly be arrived at, with low internal Na^+ and Cl^-, and high internal K^+. It appears, however, that the resting potential in nerve and muscle is a little lower than the equilibrium potential for K^+, *i.e.* there is a little more K^+ inside than could be drawn in by the resting potential. It

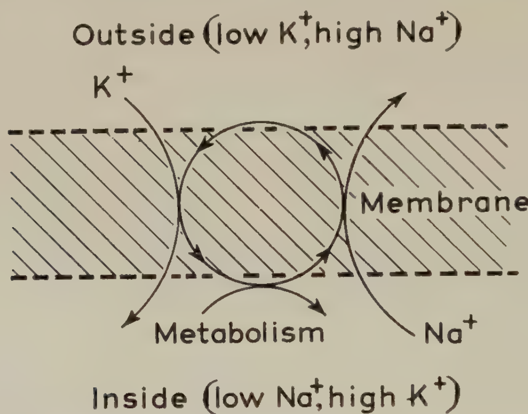


FIG. 12. 12. Diagram of the supposed mode of action of the sodium-potassium exchange pump. (After Hodgkin and Keynes.)

could only get there against the electrochemical gradient by means of a pump.

Uptake of K^+ proves to be coupled to extrusion of Na^+ in a single sodium-potassium exchange pump (Fig. 12. 12). The most direct evidence for the coupling of the two processes is that removing K^+ from the external fluid inhibits extrusion of Na^+ from inside the fibre. The only straightforward explanation of this finding is that the pump has to be able to take in K^+ in order to extrude Na^+.

The ionic differences between the cell and its surroundings which the sodium-potassium pump maintains, are the basis of the excitable properties of nervous and muscular tissues. Whether in evolution this is the reason for them is not known, but nearly all cells whether excitable or not accumulate K^+ and expel Na^+, and a fair proportion of the basal metabolism of the body is devoted to driving the pump by which they do so.

The resting permeability to Na^+ is so low that when the pump is poisoned, large nerve and muscle fibres run down so slowly that the

resting and action potentials are not altered significantly for hours. The pump is only necessary for maintaining the composition of the cells over long periods or in recovery after activity. It is important to be clear that the pump is not directly involved in the mechanism of the action potential. The concentration differences of Na^+ and K^+ across the membrane can be regarded as furnishing ionic batteries from which large brief currents can be drawn during the action potential. The pump only recharges the batteries and has nothing to do with turning on the currents during the action potential—which continues quite unaltered if the pump is poisoned.

The Action Potential. Up to the present we have spoken of the permeabilities of the nerve membrane to Na^+ and K^+ as if they had fixed values. But in fact they depend on the potential difference across the membrane and if this is altered by passing an electric current the permeabilities change. They both change in the same sense, namely that as the membrane potential is reduced the permeabilities increase, but the time relations differ. When the membrane is suddenly depolarised and held at the new value, the Na^+ permeability rises at once ; but the increase is short-lasting and within a few milliseconds the Na^+ permeability has fallen away again. The K^+ permeability, however, rises only slowly but the increase is then maintained as long as the depolarisation lasts. These facts were discovered by experiments on squid fibres in which the membrane potential was displaced by a known number of millivolts and "clamped" at the new value by an electronic device. The current flowing across the membrane, which is a measure of its permeability to ions, was recorded. If the fibre is in a special solution containing no Na^+, practically all the current is carried by K^+. Hence the record gives the change of K^+ permeability with time. With sea water outside the record gives the sum of the Na^+ and K^+ permeabilities, from which the time course of Na^+ permeability alone can be obtained by subtraction.

When an action potential is set up by an electric stimulus the sequence of events is as follows. Current leaving the fibre under the stimulating cathode causes a voltage drop to develop across the electrical resistance offered by the membrane. This voltage drop is of the opposite polarity to the resting potential and hence it causes a depolarisation of the membrane. The Na^+ permeability at once rises and Na^+ enters, driven both by the concentration gradient and the potential difference. The entry of positively charged sodium ions tends to make the inside of the fibre less negative with respect to the outside, *i.e.* it still further depolarises the membrane. This depolarisation resulting from Na^+ entry turns on still more Na^+ permeability. Thus a self-regenerative increase in Na^+ current develops, which very rapidly depolarises the membrane, and gives the rising phase of the action potential. The inward flow of Na^+ ceases when the membrane potential reaches the equilibrium potential for Na^+ with the inside of the fibre positive to the outside (see p. 363) ; this marks the peak of the action potential. By this time the Na^+ permeability is already falling, for as we have said

the increase caused by depolarisation is only transient. If nothing else happened the potential would stay near the Na^+ potential until Na^+ permeability fell to ordinary levels. The potential difference across the membrane would then be very far from the equilibrium potential for K^+, so that K^+ would leave the fibre. K^+ current would continue to flow until the inside of the fibre was sufficiently negative with respect to the outside to prevent K^+ leaving, *i.e.* until the resting potential was reached again. In fact the falling phase of the action potential begins earlier and proceeds more rapidly than this, because the delayed rise in K^+ permeability, which begins to make itself felt soon after the peak, allows much larger K^+ currents to flow and greatly accelerates the whole process. The falling phase of the action potential is thus due to an

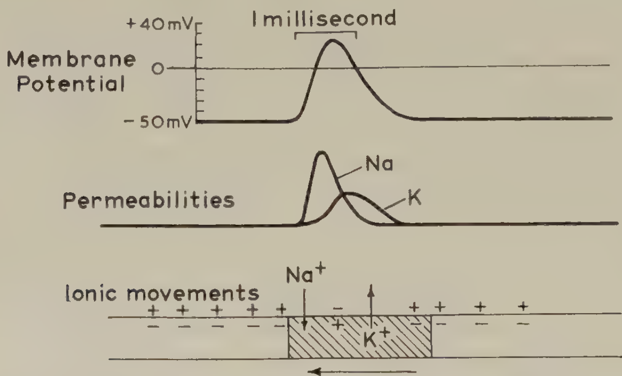


FIG. 12. 13. Diagrams showing the movements of Na^+ and K^+ during the action potential.

At the top is an action potential record, as in Fig. 12. 6. Below are shown the Na^+ and K^+ permeabilities during the action potential. The rising phase and peak of the action potential are caused by a large but transient rise in Na^+ permeability. The falling phase is due to a delayed rise in K^+ permeability.

In an action potential advancing along a fibre (bottom) Na^+ current entering at the front part of the active region reverses the potential across the membrane ; K^+ leaving the fibre in the rear re-establishes the resting potential.

outward movement of K^+. The movements of Na^+ and K^+ during the action potential are shown diagrammatically in Fig. 12. 13.

The self-regenerative increase in Na^+ permeability during the rising phase of the action potential is what gives rise to the all-or-none property of nerve. After a stimulus is applied either the membrane goes all the way to the top of the action potential or it sinks quietly back to the resting potential if the stimulus is too small. As the rising phase is self-regenerative it can never go only part of the way. The threshold, or level of depolarisation at which self-regeneration begins, is set by the resting K^+ and Cl^- permeabilities. If when the stimulus ends the Na^+ current inwards tending to depolarise is greater than the sum of the K^+ and Cl^- currents tending to repolarise, then the regenera-

tive cycle will occur, otherwise not. If the resting K^+ and Cl^- permeabilities are raised not only is the critical level of depolarisation raised but by Ohm's Law a larger stimulating current is needed to produce any given depolarisation, because the resistance of the membrane (which is inversely proportional to the sum of all the ionic permeabilities) is lowered. As we shall see, inhibition in the central nervous system is brought about by increasing the permeability of neurones to K^+ and Cl^- thereby making excitatory currents less effective.

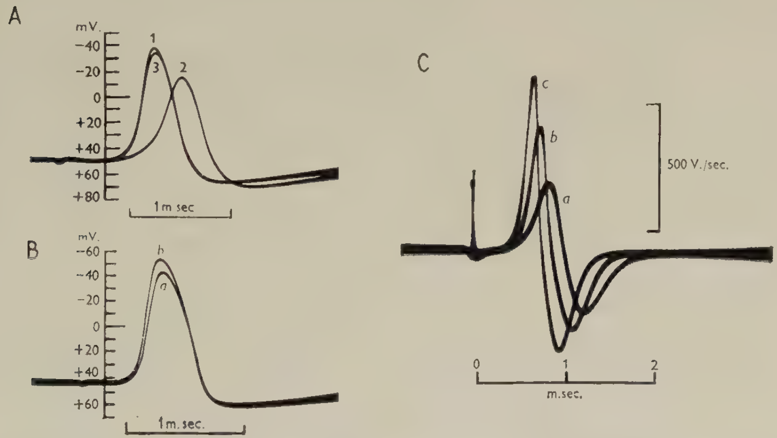


FIG. 12. 14. Effect of varying the concentration of Na^+ in the external fluid on the height and rate of rise of action potential in a single giant squid axon. An internal electrode (Fig. 12. 5) was used for recording the potential difference across the nerve membrane.

A. Sodium-deficient solution. Record 1, response in sea-water. Record 2, in a solution composed of 50 per cent. sea water, 50 per cent. isotonic dextrose (Na^+ concentration 0·5 times that of sea water). Record 3, after reapplication of sea water.

B. Sodium-rich solution. Record a, response in sea water. Record b, in sea water in which additional sodium chloride had been dissolved to make the Na^+ concentration 1·56 times that of sea water.

C. Records showing directly the rate of change of membrane potential, obtained by electronic differentiation of the ordinary action potential record. The height of the first peak in each record is proportional to the maximum rate of rise of the rising phase of the action potential (in volts/second). Record a, fibre in 50 per cent. sea water as in A. Record b, in sea water. Record c, in enriched sea water containing 1·56 times the normal Na^+ concentration as in B. (Hodgkin and Katz, *Journal of Physiology*, vol. 108, p. 37, 1949.)

The most striking piece of evidence for the sodium hypothesis is the effect of changing the Na^+ concentration in the fluid bathing the nerve. It is a direct consequence of the theory that the height and rate of rise of the action potential should vary with the external Na^+ concentration. Fig. 12. 14 shows records of action potentials with low, normal and high Na^+ concentrations. The height and rate of rise of the action potential in different Na^+ concentrations are in quantitative agreement with the hypothesis. When the Na^+ concentrations are the same inside and out-

side the nerve the action potential should just not overshoot zero potential difference, for the sodium equilibrium potential would then be zero. This is found to be so. In a sodium-free medium no action potential can be obtained at all. These results have been extended to many other tissues. Frog sciatic nerve will conduct for an hour or more after it is immersed in a sodium-free fluid ; but this is because the perineurium (the sheath around the outside of the nerve trunk) is a barrier to diffusion and prevents Na^+ getting out. Isolated nerve fibres dissected from the frog's sciatic are blocked within a second in sodium-free media. Block is also rapid in the spinal nerve roots which have no sheath.

These experiments made it highly probable that sodium ions carry the current across the membrane during the rising phase of the action potential, but the matter could not be regarded as settled until Na^+ had been shown to enter the fibre during activity. This was achieved using radioactive Na^+ ; the number of sodium ions entering during a single impulse was measured and was found to carry a large enough charge to account for the change in potential difference during the action potential.

Refractory Period. A consequence of the regenerative increase in Na^+ current during the action potential is that the nerve cannot be excited again until some time after the regenerative cycle is finished. Obviously a second regenerative process cannot be superimposed on the rising phase. After the peak, Na^+ permeability is low (" inactivated " is the expression used) and only recovers as the nerve repolarises. The regenerative process can be made to occur again when inactivation has diminished sufficiently for it to be possible by depolarising to obtain a Na^+ current larger than the sum of the K^+ and Cl^- currents tending to repolarise. A larger stimulus than usual will be necessary until the nerve returns to its resting condition for three reasons ; firstly inactivation makes it more difficult to turn on the Na^+ current, secondly until K^+ permeability returns to normal a larger Na^+ current will be needed to exceed it and cause regenerative action, thirdly by Ohm's law a larger current than usual will be needed to depolarise by a given amount because the membrane resistance is lowered so long as the K^+ permeability is raised.

The practical effect of these events is that during the spike no stimulus will excite ; this is the " absolute refractory period " lasting about 1 millisecond in frog nerve. Then for about another 5 milliseconds (the " relative refractory period ") a larger stimulus than usual is necessary. During the relative refractory period the same factors that raise the threshold also decrease the rate of rise of the second spike, reduce its amplitude and cause it to be conducted more slowly.

The refractory period is important because it sets an upper limit to the frequency at which nerve fibres can carry impulses. The limit is about 800 per second in the frog and perhaps 1,600 per second in mammals. The frequencies met with in life are usually much lower than this, probably because the above values only apply to the first few

impulses. If a prolonged train of impulses is set up cumulative factors enter and later impulses require much larger stimuli. Hence for a sustained discharge only much lower rates are possible.

Accommodation. If a subthreshold current is applied to a nerve the resulting depolarisation, although insufficient to set off a regenerative change, does result in some inactivation of the Na^+ permeability mechanism and turns on the delayed rise in K^+ permeability. The threshold therefore rises for exactly the same reasons that it is raised during the relative refractory period ; these changes are thought to be the cause of accommodation—the process which makes the nerve inexcitable by slowly rising stimuli (p. 348).

It has been known for a long time that excitable tissues lose their accommodation if placed in solutions without calcium ; the threshold falls and small constant currents will cause the discharge of a continuous series of impulses. Eventually impulses arise spontaneously. Calcium ions have been shown to have a profound effect on the permeability mechanisms for sodium and potassium in squid and frog nerve.

Clinically the failure of accommodation in motor fibres is the cause of the spontaneous contractions of muscles, particularly those of the hand and foot (carpo-pedal spasms), that occur in states of low blood calcium ; for instance when the parathyroid glands are damaged accidentally in the course of an operation on the thyroid. Accommodation also fails in nerves recovering from anoxia. This is the cause of the spontaneous discharge of sensory nerves called " pins and needles " when a tourniquet is removed from a limb, or when the legs are uncrossed after sitting still for some minutes. In the latter case the peroneal (lateral popliteal) nerve is deprived of its blood supply by pressure where it passes round the head of the fibula.

Accommodation is not equally rapid everywhere in an axon. In sensory nerves it is slow near the sensory end organs, so that slowly rising generator potentials arising in the end-organ give rise to trains of nerve impulses. Similarly it must be slow or absent at the origin of the motor fibres in the ventral horn, because a steady depolarisation of the motoneurone cell body causes a maintained discharge of impulses in the motor fibre. The basis of these differences in accommodation at the ends of axons has not yet been investigated in terms of the sodium and potassium permeability mechanisms.

Saltatory Conduction in Myelinated Nerves

We have already seen that increasing the strength of the local currents spreading in front of the action potential by increasing the diameter of the fibre, increases the conduction velocity (p. 359). This is why invertebrates have developed giant fibres. The vertebrates obtain much higher velocities with smaller fibres by another method, making use of the property that local circuits can be made to spread further ahead of the active region by increasing the electrical resistance of the membrane. This can easily be understood by analogy with a submarine telegraph cable ; if the insulation is bad the signal will not reach the

other end but will be lost in local circuits through the insulation ; the better the insulation the further the signal will travel. Vertebrates improve the insulation by laying down thick layers of myelin around the fibres, only leaving the membrane exposed at the nodes of Ranvier, spaced roughly every millimetre along the fibre. The action potential is confined to the nodes, the part of the fibre covered in myelin being inexcitable. A high conduction velocity is obtained because the currents produced by the active region, instead of being used to excite the section of nerve immediately ahead, are used to excite the next node some 50 fibre diameters further on. This is possible only because the insulation conferred by the myelin sheath ensures that the local currents get that far and do not leak away through the membrane first. Myelinated fibres transmit about ten times as fast as non-myelinated fibres of the same outside diameter at the same temperature. A typical figure is 90 metres per second for a 15 μ myelinated fibre at 37° C.

Evidence that the myelin is an insulator and that activity only occurs at the nodes has been obtained using single fibres dissected from frog's nerves, a technique that has only been mastered by a handful of physiologists. It has been shown that the threshold is much lower when the stimulating cathode is opposite a node, and that, during the rising phase of the action potential, inward current (*i.e.* sodium current) only flows through the nodes. By recording from one node after another it was shown that the action potential jumps from one node to the next with a brief delay in between, hence the description " saltatory " from the Latin *saltare*, to leap. The saltatory theory met with opposition from those who believed that there were no nodes in the myelinated nerve fibres of the central nervous system. Ranvier himself said there were none, but they were seen by Cajal and their presence has been amply confirmed by recent investigations.

Another advantage of saltatory conduction, apart from the increased speed, is that the amounts of sodium gained and potassium lost during activity are much decreased, owing to the very much smaller area of active membrane. The amount of sodium entering per impulse is about 300 times smaller in frog's nerve than in an unmyelinated fibre of the same size. Hence the metabolic energy needed for recovery is much less and the fibres are practically indefatigable.

Nerve Fibre Diameter in relation to Function

Under the microscope a vertebrate nerve trunk proves to contain myelinated nerve fibres of all sizes from roughly 20 μ down to 1 μ in diameter. There are also numerous small non-myelinated fibres, mostly under 1 μ in diameter, which in many nerves outnumber the myelinated fibres. The relative numbers of myelinated fibres of different diameters varies from nerve to nerve, and between cutaneous and muscular nerves. The fibre spectrum of a cutaneous nerve from a cat is shown in Fig. 12. 15.

As we have already seen on p. 359, large nerve fibres conduct faster than small. They also have a lower threshold and give rise to

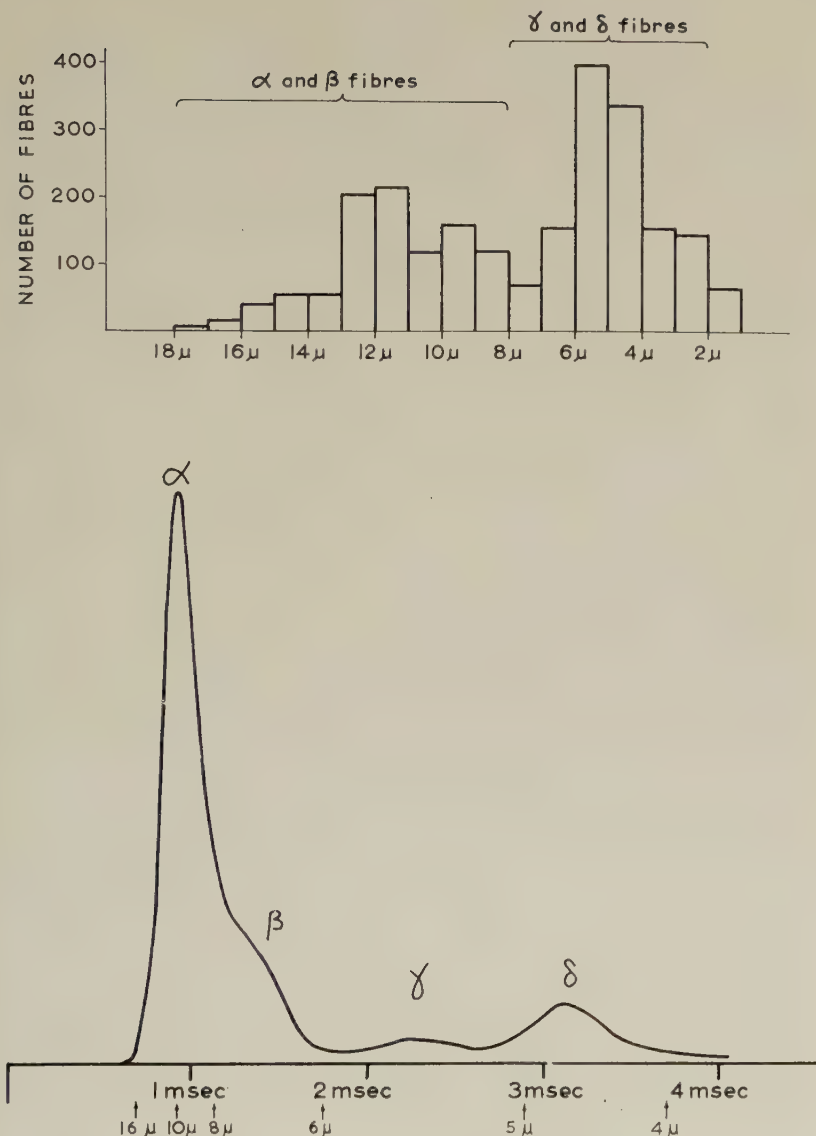


FIG. 12. 15. Fibre spectrum and compound action potential of a cat's saphenous nerve. The graph shows the numbers of myelinated nerve fibres of different diameters in the saphenous nerve (a cutaneous nerve) of the cat.

Below is a tracing of the action potential obtained from the same nerve. The conduction distance was 6 cm. The diameters of the fibres responsible for the various parts of the compound action potential are given below. (Modified from Gasser and Grundfest, *American Journal of Physiology*, vol. 127, p. 393, 1939.)

larger action potentials. Both these properties, like their higher conduction velocity, depend on the lower longitudinal resistance of the larger axons. The voltage change across the membrane during the action potential (which is set by the Na^+ and K^+ concentrations and has the same amplitude and time duration in all the myelinated fibres) gives rise to larger currents in the external fluid when the axon resistance is low, and hence to a larger potential recorded with electrodes outside the nerve. Conversely, with the larger axons, a smaller applied voltage is necessary in order to cause a current of threshold strength to flow in them. As a result of these properties, if we take a length of nerve, stimulate it at one end and record the action potential from the other, the action potential just above threshold arrives with the shortest latency and is a simple spike in form ; it corresponds to a volley in the largest fibres. A larger shock brings in the next slower fibres which arrive later and produce a hump on the descending phase of the original spike ; and so on with larger and larger shocks.

The action potential of the cat's saphenous nerve, with a stimulus large enough to excite all the myelinated fibres, is shown in Fig. 12. 15. In the terminology introduced by Erlanger and Gasser, who first investigated the properties of nerve fibres of different sizes in the nineteen-twenties, the myelinated fibres of peripheral nerve belong to the A group. The A group action potential is subdivided into α, β, γ and δ elevations, as indicated in the figure. Because of their smaller action potentials the γ and δ fibres, although more numerous than the α and β fibres, make a much smaller showing in the compound action potential. The conduction velocities of the α and β fibres run from 90 metres/sec. down to, say, 45 metres/sec. The γ and δ fibres go from there down to 2 or 3 metres/sec. for the fibres of about 1 μ ; but in the record in Fig. 12. 15 the contribution from the smallest fibres is not seen.

The non-myelinated fibres are termed the C group. Their diameters are roughly 0·4 to 1·0 μ, with conduction velocities of about 1 metre/sec. If the record in Fig. 12. 15 had been taken on a much slower time scale the main C fibre action potential would appear as a very small elevation at about 60 msec.

In cutaneous nerves there are fibres of every size that respond to mechanical stimulation of the skin. The lightest touches (of the kind that evoke sensations of itch and tickle in man) give rise to impulses in δ and C fibres only ; firm stimulation brings in α fibres. Painful stimuli appear to be signalled both by slow conducting A fibres, and by C fibres. Warming and cooling the skin also cause impulses in the slower A fibres and in C fibres.

In muscular nerves the motor fibres that cause the muscle to contract belong to the α group. There are also γ motor fibres that run to supply the small muscle fibres in the muscle spindles—one of the types of sensory ending in muscle (see p. 411). The sensory endings in muscle are connected to both fast and slow conducting A fibres.

The fine myelinated preganglionic fibres in the white sympathetic rami, and elsewhere in the autonomic nervous system (Chapter 15),

are put in a special group, the B group. They are similar in size and conduction velocity to δ fibres of the A group. The non-myelinated postganglionic sympathetic fibres belong to the C group.

Fast conducting nerve fibres take up more space and use up more metabolic energy than small ones. This must be one reason why not all nerve fibres are large. It seems clear enough why some touch fibres and motor fibres have to be large, and why fibres conveying ill-localised aching pain can afford to conduct slowly. But much remains to be discovered particularly about the allocation of different modalities of sensation to the different groups and the reasons for it.

Nerve Block

Conduction in nerve fibres is blocked by many substances with a general anæsthetic action, such as alcohol, morphine, and ether or chloroform vapour. More powerful and convenient than these is cocaine, and its safer derivatives, procaine, etc., dilute solutions of which are widely used in medicine for local anæsthesia. Cocaine blocks small nerve fibres before large ; hence the well-known paradox that cocainised surfaces may be insensitive to pain and yet able to feel the brush of a wisp of cotton wool (see also p. 414).

Nerves also cease to conduct if deprived of oxygen, but anoxia blocks the myelinated A fibres before the smaller unmyelinated C fibres. Hence in human limbs deprived of blood (ischæmic), touch and motor power fail before sensitivity to pain. When a tourniquet, or, better, the pneumatic cuff of a blood-pressure measuring machine (sphygmomanometer) inflated to a pressure above systolic, is applied to the upper arm, the first symptoms of block are numbness of the finger tips coming on after about 14 minutes followed by loss of touch sensation in the fingers and hand, and paralysis of the small muscles of the hand. Subsequently anæsthesia and paralysis slowly ascend the arm. After an hour the whole arm from the cuff down is anæsthetic to touch and paralysed ; but pin pricks are still perceived as painful, although the sensation is delayed by a fraction of a second. This persisting, delayed pain is due to unblocked, slowly conducting C fibres. Ischæmia for longer than an hour is not advised (although periods of $1\frac{3}{4}$ hours have been used for experimental purposes, so far with impunity) because of uncertainty as to the time of onset of the disastrous irreversible changes in ischæmic muscles, akin to rigor mortis, occasionally seen in limbs constricted by splinting or plaster. When the circulation is restored after occlusion for an hour or less, recovery of sensation and movement begins within half a minute and proceeds rapidly. The familiar " pins and needles " follows (see p. 371).

It was proved by Thomas Lewis and his co-workers that the cuff blocks conduction because it cuts off the blood supply to the nerve, and not because of the pressure it exerts on the nerve fibres. Thus provided the pressure is always above systolic blood pressure, block develops in the same time whether the cuff be inflated to 150 or to 300 mm. Hg. Again if a cuff be applied just above the elbow until, say, the hand is

anæsthetic to touch, and then a second cuff be inflated on the upper arm above the first and the first cuff removed, there is no recovery in the hand. The pressure has been removed from the original length of nerve without recovery ; hence it cannot be the pressure that is relevant to blocking, but the ischæmia that is maintained by the second cuff.

It was also shown that anæsthesia and paralysis are due to conduction block in the proximal part of the nerve, and not to the effect of ischæmia on sensory endings and on the muscles. Thus, if one waits until the hand is anæsthetic and its intrinsic muscles paralysed, and then inflates a cuff round the wrist, subsequent removal of the cuff on the upper arm results in full recovery of sensation and power in the hand, although the hand remains ischæmic. Clearly the block lay in the stretch of nerve between the upper cuff and the wrist. The reason that the proximal part of a nerve fibre is more susceptible to ischæmia than the distal part is not known ; neither is it known why the longest fibres suffer first so that symptoms appear first in the finger tips.

CHAPTER 13

NEUROMUSCULAR AND SYNAPTIC TRANSMISSION

Neuromuscular Transmission

THE evidence that propagation of the impulse along nerve fibres is electrical is now overwhelming and its mechanism has been described. What happens when an impulse reaches the end plate where the nerve terminates on the surface of a muscle fibre? After a delay of about a millisecond an action potential, similar in nature to that in nerve, arises in the neighbourhood of the end plate and propagates by local circuits in the ordinary way along the muscle fibre membrane in both directions. How does the nerve impulse set off a muscle action potential? The arrival of the active region, which is negative with respect to the rest of the nerve, is equivalent to applying a weak cathodal stimulus to the muscle fibre at the end plate. This will tend to depolarise and excite the muscle fibre, but does it succeed? The answer is, no. Quantitatively the electrical stimulus is not strong enough. The nerve fibre does not supply nearly sufficient current to depolarise the very much larger area of the muscle fibre membrane. Instead, when the impulse arrives, the nerve ending secretes a very small quantity of the chemical substance acetylcholine. Acetylcholine has practically no effect on nerve or on most of the muscle fibre, but at the end plate, opposite the nerve ending, the muscle fibre membrane is highly sensitive to acetylcholine, to which it responds by an immediate and very large increase in permeability to both sodium and potassium ions. The effect is equivalent to making a minute hole in the membrane through which ions can pass freely and, as with a mechanical puncture, the resting potential disappears at that point. This depolarisation of the end plate has the same action as applying a strong cathodal electrical stimulus to the muscle membrane around it, which thereupon is rapidly depolarised and excited. Within a millisecond or two the acetylcholine is destroyed by an enzyme, choline-esterase, present in high concentration in the end plate, which hydrolyses it into inactive acetic acid and choline. Thus acetylcholine and the nerve action potential itself both act in a manner which would tend to depolarise the muscle fibre. The difference is that in practice any depolarisation due to the nerve action potential is too small to be detected, whereas that due to the acetylcholine excites by a comfortable margin. The chemical mechanism can be regarded as a device for amplifying the electrical effect of the nerve impulse by making a temporary chemical puncture in the muscle membrane.

Conclusive evidence for chemical transmission by acetylcholine has only recently been obtained. The story goes back to Claude Bernard's experiments with the South American arrow poison, curare,

in 1850. Curare paralyses striated muscles. Bernard showed that it does so by blocking the passage of the nervous impulse from nerve to muscle. After curare a nerve-muscle preparation no longer contracts when the nerve is stimulated, but it does so again if the stimulating electrodes are moved on to the muscle itself ; hence curare does not act by making muscle fibres inexcitable. Nor does it act on nerve fibres, for after painting curare on the nerve alone the muscle contracts in the ordinary way when the nerve is stimulated. Therefore there must be a region with special properties between nerve and muscle, now identified as the end plate, where curare has its effect. Curare was later found to act by rendering the muscle membrane at the end plate insensitive to acetylcholine ; acetylcholine is still released by the nerve impulse but no depolarisation of the muscle membrane follows. For a hundred years curare remained a physiological curiosity and tool, but recently the purified active constituent, curarine, and other drugs with analogous actions have come into extensive use in surgery for assisting muscular relaxation at operation and thereby permitting a lighter general anæsthesia.

At the beginning of this century Langley showed that the neuromuscular junction is specially susceptible to chemical excitation as well as to block. In most muscles the end plates lie near the centre of the fibres and in parallel-fibred muscles they form a band across the middle of the muscle near the point where the nerve enters. When dilute nicotine solution is applied locally to the end plate region prolonged twitching of the muscle results, but not when it is put on the nerve-free parts of the muscle or on the nerve itself. The chemically excitable structure revealed by nicotine appears to be the structure that the nerve impulse normally excites, for like the nerve impulse, nicotine action is blocked by curare.

These results would naturally suggest that the nerve fibre itself might excite the muscle by means of a chemical substance liberated at the nerve ending. As a matter of fact the idea of chemical transmitters had already arisen in connection with the autonomic nervous system (Chapter 15). It was known that injection of adrenaline, a substance isolated chemically from the adrenal glands, into the blood, gave rise to many of the effects of sympathetic nerve stimulation, cardiac acceleration, rise of blood pressure, etc. Similarly, injection of muscarine, the poison of a common toadstool, mimicked parasympathetic nerve action, slowing of the heart, salivation, etc.

Definite proof that the nerve impulse does liberate a chemical transmitter was first achieved by Otto Loewi in 1921. He showed that when the isolated frog's heart is slowed by vagal stimulation the perfusion fluid that comes from it will cause slowing of a second heart (Fig. 15. 1, p. 462). The transmitter secreted by vagal endings was identified by Dale and his colleagues as acetylcholine. Refined pharmacological tests had to be developed, for the amounts involved were far too small for chemical methods (Fig. 15. 3, p. 465). Dale and his colleagues afterwards applied similar methods to detect the even more

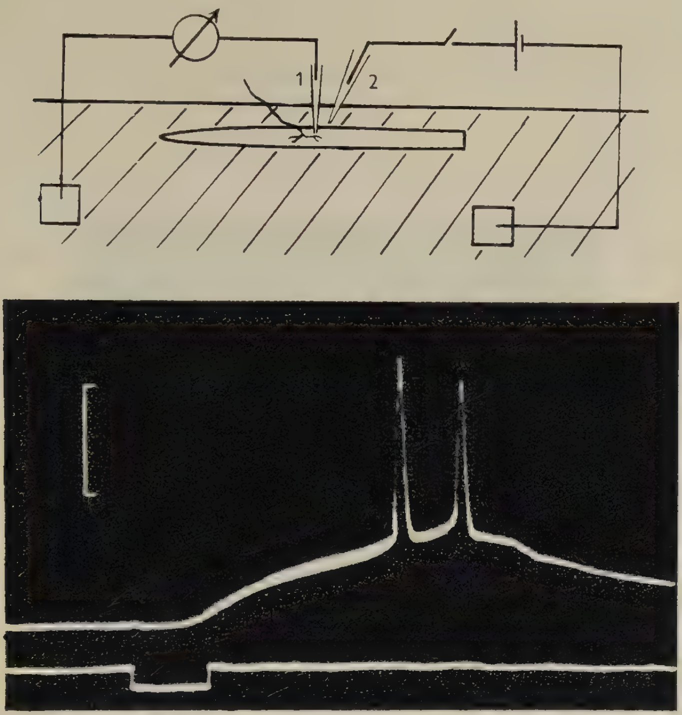


FIG. 13. 1. Effect of applying acetylcholine to a motor end-plate in frog muscle.

The diagram shows the experimental arrangements. A glass micropipette (1), containing potassium chloride solution, is inserted into a muscle fibre at the end-plate region and used to record the membrane potential. A second micropipette (2) containing a solution of acetylcholine, is manœuvred as close as possible to the outside of the end-plate ; an outwardly directed pulse of current ejects acetylcholine from the tip of this electrode by electrophoresis. The electronic equipment used to produce brief current pulses is represented in the diagram by a battery in series with a make-and-break switch.

In the record, the step-like deflection in the lower trace signals the current pulse through the acetylcholine pipette. After the pulse, the potential record (upper trace) shows a slow depolarisation of the end-plate membrane (an *end-plate potential*) on which are superimposed two propagated action potentials.

Vertical calibration line, 50 mV. The current pulse lasts 17 milliseconds. (Katz, *Bulletin of the Johns Hopkins Hospital*, vol. 102, p. 275, 1958.)

minute amounts of acetylcholine in the blood coming from stimulated skeletal muscle ; for this pharmacological *tour de force* the choline esterase in the muscle, which otherwise would have destroyed the acetylcholine immediately after release, was inhibited by the drug eserine.

These and other experiments made it clear that acetylcholine acts as the transmitter of excitation across the neuromuscular junction. The details of how it works have now been unravelled by modern microelectrode techniques, mainly in the hands of Katz and his co-

workers. They employed a method of applying small doses of acetylcholine direct to the motor end plate by electrophoresis from a micropipette containing the drug. Another similar micropipette filled with KCl solution was inserted into the muscle fibre nearby and used as an electrode to record the potential difference between the inside and outside of the end plate membrane (Fig. 13. 1). A squirt of acetylcholine,

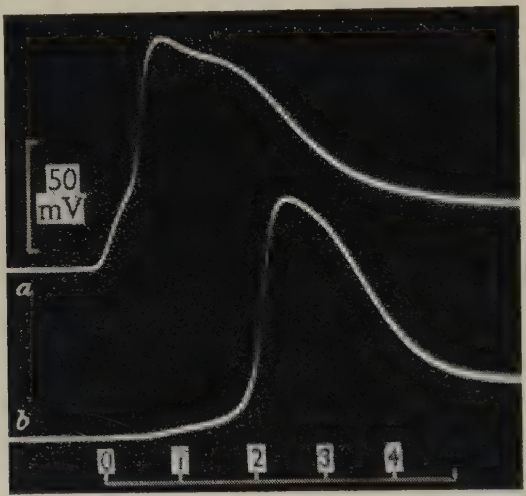


Fig. 13. 2. Neuromuscular transmission in frog muscle recorded with an intracellular microelectrode.

A glass microcapillary is used to record the membrane potential of a single muscle fibre as in Fig. 13. 1. The records show what happens after a nerve impulse, set up by stimulating the motor nerve at a distance from the muscle, arrives at the muscle end-plate.

Record *a*, with the microelectrode in the end-plate region, shows a rapidly rising end-plate potential from which an action potential takes off. The point at which the action potential arises is marked by a " step " on the rising phase. (A more conspicuous " step " is to be seen in Fig. 13. 3, A). End-plate action continues during the action potential and causes a " hump " on its falling phase.

In record *b* the microelectrode in the same muscle fibre, 2·5 mm. away from the end-plate, picks up the propagated muscle action potential travelling away from the end-plate. The longer latency before the action potential corresponds to the time taken for conduction from the end-plate region. At this distance there is little sign of the end-plate potential.

Such records show, first that the end-plate potential is a local non-propagated response and, secondly that the propagated muscle action potential arises from the end-plate region.

Time scale in milliseconds. (Fatt and Katz, *Journal of Physiology*, vol. 115, p. 320, 1951.)

delivered by passing a brief pulse of current through the acetylcholine pipette, gives rise to a depolarisation of the end plate, the end plate potential, which, if it is large enough, triggers off a muscle action potential (Fig. 13. 1). Acetylcholine has no effect if it is applied to the inside of the end plate membrane after the micropipette has been pushed through the membrane into the muscle fibre. With the pipette outside the effect falls off rapidly if the pipette is moved along the fibre away

from the end plate. Even with the most careful positioning the pipette cannot be got as close to the muscle membrane as the nerve terminals, which lie in little troughs of membrane on the surface of the fibre, so that a larger quantity of acetylcholine is necessary than a nerve impulse secretes, and the concentration at the membrane rises and falls more slowly. The end plate potential due to a nerve impulse is a much briefer affair and the depolarisation is so rapid that it may be difficult to see where the muscle action potential takes off (Fig. 13. 2).

It is not known how acetylcholine causes the greatly increased

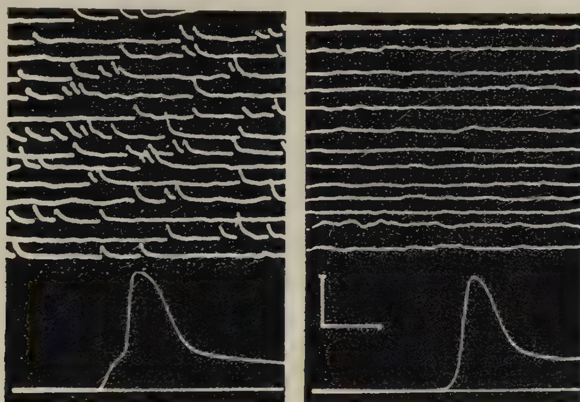


Fig. 13. 3. Spontaneous miniature end-plate potentials in frog muscle.

An intracellular microelectrode was used to record the membrane potential of a single muscle fibre as in Figs. 13. 1 and 13. 2.

In A, the electrode was at the end-plate region. The upper part consists of a number of records taken at slow speed and high amplification, showing the small depolarisations (miniature end-plate potentials) which go on irregularly all the time in the absence of stimulation. In B, the electrode was inside the same muscle fibre 2 mm. away from the end-plate. No miniature end-plate potentials are seen.

The lower parts of A and B show the response to nerve stimulation, recorded at a lower gain and faster sweep from the same sites. In A there is a conspicuous end-plate potential from which the action potential takes off. No end-plate potential is visible in B. These records, therefore, confirm that the microelectrode was at the end-plate in A (compare Fig. 13. 2).

Voltage and time calibrations are given by the L-shaped scales in B. For the upper records the scales represent 3·6 mV. and 47 msec. ; for the lower records, 50 mV. and 2 msec. (Fatt and Katz, *Journal of Physiology*, vol. 117, p. 109, 1952.)

ionic permeability which depolarises the end plate. The changes in permeability to sodium and potassium ions are certainly quite different from those that occur during an action potential. The basis of the action potential is first that the permeabilities to sodium and potassium ions only alter when the potential difference across the membrane alters and secondly, that these changes are separated in time, the brief rise in sodium ion permeability preceding the rise in potassium ion permeability. At the end plate, however, the permeability changes

are simultaneous and independent of membrane potential. Thus if a muscle is placed in an isotonic solution of a potassium salt, the resting potential vanishes because the concentration of potassium ions is the same inside and outside. In this condition the muscle and nerve become completely inexcitable, but electrophoretic application of acetylcholine to the end plate although it causes no potential change still results in a large increase in permeability, as evidenced by a fall in the electrical resistance of the membrane. The membrane resistance is measured by passing a known current through the membrane from an internal micropipette, observing the alteration in membrane potential produced and applying Ohm's law.

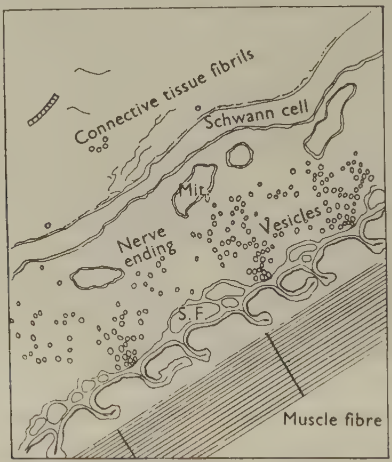


FIG. 13. 4A. Outline drawing showing the principal structures visible in the electron micrograph, Fig. 13. 4B.

The picture shows a fine terminal filament of the motor nerve fibre lying in contact with the surface of the muscle fibre. The area of contact is increased by folding of the muscle fibre membrane at intervals of some 0·4 μ. A Schwann cell covers the outside of the nerve fibre, and fingers of the Schwann cell (S.F.) protrude into the space between nerve ending and muscle fibre. The nerve ending contains vesicles (thought to contain acetylcholine) and mitochondria (Mit.)

(Birks, Huxley and Katz. *Journal of Physiology*, vol. 150, p. 134, 1960.)

Quantal Release of Transmitter. When internal microelectrodes were first used to study electrical events at the end plate it was discovered that, even when no nerve impulses are arriving, the end plate is not electrically silent : small spontaneous potential changes are recorded from time to time as shown in Fig. 13. 3 ; they are similar in shape to ordinary end plate potentials but only about 1 per cent. of the size (0·5 mV. instead of some 50 mV.). An obvious explanation would be that these miniature end plate potentials are due to accidental leakages of single molecules of acetylcholine from the nerve terminal. They appear, however, to be very much too large for that. For this and other reasons it is believed that they represent the simultaneous

release of a very large number of acetylcholine molecules in a packet. This is certainly a surprising notion but electron microscopy has revealed that the nerve terminals contain numerous "vesicles," to be seen in Fig. 13. 4, which are strongly suspected to contain the packets of

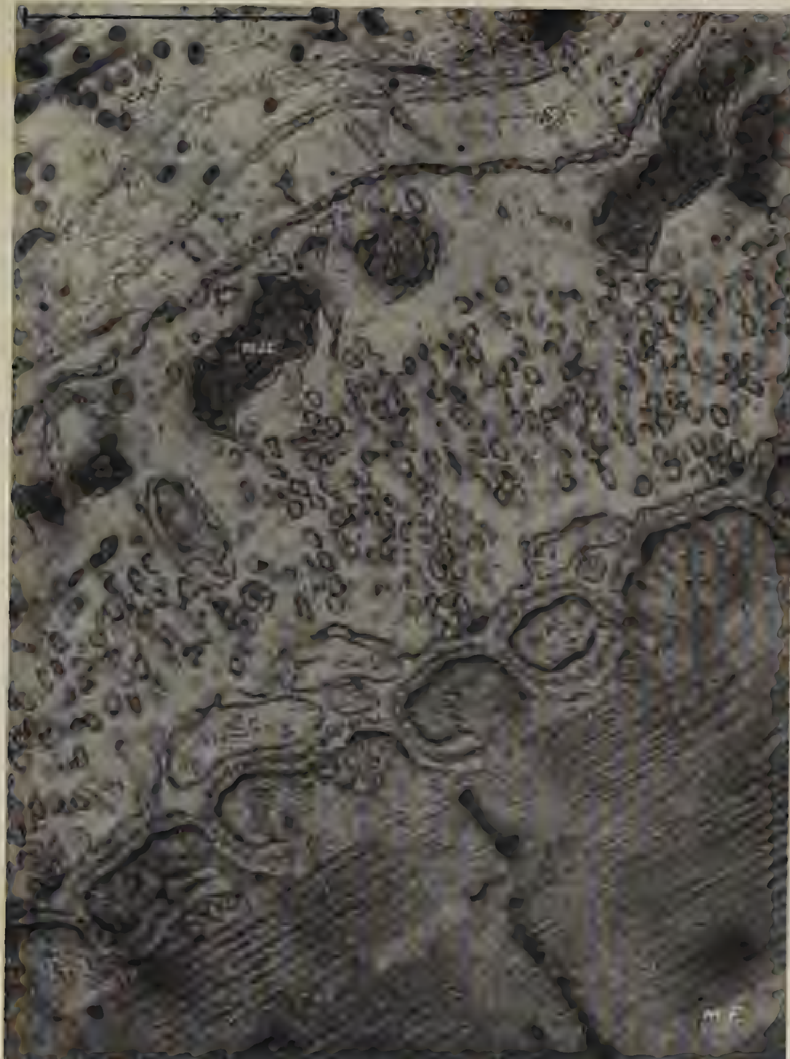


FIG. 13. 4a. Electron micrograph of the neuro-muscular junction of a frog's sartorius muscle. Longitudinal section of the muscle. The scale at the top is 1μ long.

acetylcholine in question. Pharmacological tests have shown that motor nerve fibres contain acetylcholine and also an enzyme *choline acetylase* which synthesises it.

How the packets of acetylcholine get out is not known but the

rate at which they do so is found to depend on the potential difference across the nerve terminal membrane. If the nerve terminal is depolarised by passing a small direct current through it the rate of occurrence of miniature potentials rises steeply. Whatever the rate, the size of the miniature potentials remains the same ; acetylcholine appears always to be released in packets of the same size. When a nerve impulse arrives, it causes a large and rapid depolarisation of the terminal which for a millisecond or so raises the rate of release of acetylcholine packets so high that a hundred or two miniature potentials occur on top of each other. They add up to give a single large end plate potential.

Hypersensitivity of Denervated Muscle. When the motor nerve to a muscle is cut and allowed to degenerate the muscle becomes hypersensitive to acetylcholine and gives a larger and more prolonged contraction than usual when a small quantity of acetylcholine is injected into the artery supplying the muscle. This state is not permanent for eventually the muscle fibres atrophy if the motor nerve does not regenerate. Electrophoretic application of acetylcholine to hypersensitive muscle has shown that the phenomenon is not due to an increased sensitivity of the end plate but to an extension of acetylcholine sensitivity over the whole muscle fibre. Small quantities of acetylcholine in the fluid around the fibre have a greater effect than usual because the area that is sensitive is enormously increased, not because any one area has become more sensitive. In an embryo the muscle fibres are at first sensitive to acetylcholine all over, but when motor fibres grow out and make connection with them the sensitivity to acetylcholine shrinks on to the end plate. The implications of these extremely interesting discoveries have yet to be worked out. Skeletal muscle is not the only tissue that becomes hypersensitive when it is denervated. For instance, after cutting the sympathetic nerves to a limb the blood vessels become hypersensitive to circulating adrenaline.

Neuromuscular Block. When a prolonged train of nerve impulses arrives at a neuromuscular junction the end plate potential set up by each impulse gradually gets smaller, until it may become inadequate to excite the muscle fibre. The decrease in size of the end plate potential is primarily due to a decrease in the amount of acetylcholine liberated by each impulse. In healthy subjects there is good evidence that at the frequencies met with during prolonged reflex or voluntary activation of muscle, say 20–40 impulses per second, neuromuscular block does not occur, at least until the contractile mechanism is itself almost exhausted. Fatigue in ordinary life is, therefore, not due to neuromuscular block, but to failure of the muscle fibres to contract when an action potential passes along them. Electrical excitation of motor nerves at higher frequencies may, however, induce neuromuscular block before the contractile mechanism is exhausted.

Neuromuscular block develops much more quickly in isolated frog muscle than in mammalian muscle with intact blood supply, hence neuromuscular block can readily be demonstrated in fatigued frog muscle in the class room.

Synaptic Transmission

The Neurone Theory. Neuromuscular transmission is only a special case of the general process of the transmission of nervous activity from one excitable cell to another. In the nervous system, as elsewhere, the living units are cells, here called *neurones*, each a mass of cytoplasm bounded by a cell membrane and possessing a nucleus. When any part of the cell is separated from the nucleus it dies. The cell theory applied to the nervous system is called the *neurone theory.* Nerve fibres are processes of neurones and when separated from the cell body, containing the nucleus, *e.g.* by section of a peripheral nerve, they die and disintegrate. The process is called Wallerian degeneration after Waller who first described it in 1850.

For other tissues the cell theory was accepted as soon as the microscope and microscopical staining techniques were sufficiently developed to show clearly the individual cells and the membrane surrounding them. In the nervous system, however, the cell processes may be several feet long, so that it is not possible in general to see the whole of a single cell and be satisfied that it is marked off by a cell membrane from all other cells. As a result the neurone theory was still in dispute at the beginning of this century. Modern electron microscopy has confirmed that the cell membrane is continuous over the surface of the neurone.

Synaptic Action on Motoneurones. The problem of how activity in one neurone influences activity in another, therefore appears to boil down to the question of how electrical changes at the surface of one neurone can cause electrical changes in another. Such interaction is thought to occur largely, but perhaps not exclusively, at special sites where processes of two neurones come almost into contact with each other. Such a site is called a *synapse*, from the Greek word meaning contact. In the central nervous system synaptic action has chiefly been studied in the neurones of the spinal cord that give rise to the motor nerve fibres to skeletal muscle, the motoneurones (Fig. 12. 1, p. 346). The cell bodies and dendrites of the motoneurones are covered with synaptic endings, which take the form of small swellings at the ends of fine nerve fibres, closely applied to the motoneurone membrane. The source of some of the nerve fibres which end in this way is known. As further described in the next chapter (p. 411) muscles contain receptors responding to stretch of the muscle, impulses from which run in sensory nerve fibres to the spinal cord where they reflexly excite the motoneurones of the muscles from which they come. These sensory fibres divide many times inside the cord and some of the numerous fine nerve filaments resulting end by forming synapses on the motoneurones.

The mode of action of these particular synapses appears to be closely similar to that of the neuromuscular junction, according to the evidence obtained by recording the membrane potential of the motoneurone with an internal microelectrode. This technical feat was first achieved by

Eccles and his co-workers in 1951. Using anæsthetised cats they first removed the back muscles and the bony roof of the spinal canal to expose the spinal cord ; a capillary microelectrode filled with potassium chloride solution was then pushed through the spinal cord from the dorsal surface blindly into the ventral horn where the motoneurones lie. The motoneurones are among the largest nerve cells in the mammalian central nervous system, with a diameter of some 70 μ, and they offer a reasonable target. Impalement with a microelectrode of tip diameter about 0·5 μ does not apparently damage them, for they show a steady resting potential of roughly 70 mV. When one of the smaller neurones of the grey matter is entered, however, the resting potential is apt to decline rapidly. A cell is identified as a motoneurone by stimulating the ventral root of that segment. This sends a volley of impulses backwards (antidromically) up the motor nerve fibres into the ventral horn and if the electrode is in a motoneurone an action potential is recorded. The action potential of the motoneurone overshoots zero potential and in general appears to be similar in nature to the action potential in the peripheral motor fibre that arises from it.

When a volley of impulses in the sensory nerves from a muscle arrives at the synaptic endings, the motoneurone suffers a rapid depolarisation which lasts for some 10 milliseconds. This is called a post-synaptic potential ; it looks like and is similar in mechanism to an end plate potential in muscle. Thus by passing current through the motoneurone membrane from an internal electrode and measuring the change in potential difference produced, it is found that the membrane resistance falls steeply at the start of the post-synaptic potential. Other evidence suggests that this fall in resistance and the depolarisation that accompanies it are both due to an increase in ionic permeability to sodium, potassium and probably other ions, similar to that which occurs at the end plate.

It is presumed that these permeability changes are due to a similar type of chemical puncture of the membrane under the synaptic endings by a chemical transmitter released from them when a nerve impulse arrives. Electron microscopy reveals numerous " synaptic vesicles " in the synaptic knobs, very similar in appearance to the vesicles in the nerve terminals at the end plate, which also are suspected to be packets of transmitter. Its chemical nature is not known except that it is not acetylcholine. Acetylcholine is ruled out on strong indirect evidence. It appears that a nerve fibre which liberates acetylcholine at its terminals, a *cholinergic* nerve in Dale's terminology (p. 466), is cholinergic throughout. Acetylcholine and choline acetylase, the enzyme that synthesises it, can be detected not only at the terminals but all along the nerve fibre. Thus ventral roots contain acetylcholine and choline acetylase. But dorsal roots contain neither, and since dorsal roots contain large numbers of the fibres that run from muscle receptors to excite motoneurones it is clear that these fibres are not cholinergic.

Although the details of the depolarisation under a single nerve terminal are very similar in both the motoneurone and the end plate,

synaptic transmission in the spinal cord is otherwise arranged very differently from the neuromuscular junction. In mammals each muscle fibre bears only one or two end plates supplied by a branch or branches of a single motor nerve fibre. Except in extreme fatigue every motor nerve impulse gives rise without fail to a muscle action potential, for each nerve impulse causes an end plate potential several times larger than is needed to excite the muscle membrane. As the all-or-none law implies, the nervous system can only alter the strength of contraction in a muscle fibre by altering the rate at which nerve impulses reach it. Gradation of response occurs because although each nerve impulse gives rise to a similar all-or-none action potential in the muscle fibre, the contractile mechanism activated by the muscle action potential is not all-or-none ; it is a graded process. A single muscle action potential results in a partial and long lasting contraction ; a following action potential causes a further contraction which, if it begins within a certain interval, summates with the first. The degree of summation is determined by the frequency of muscle action potentials and thus frequency determines strength of contraction. In the motoneurone strength of excitation is also related to frequency of arrival of nerve impulses at the synaptic endings ; but the graded, non-all-or-none stage where summation takes place is the membrane potential of the motoneurone. In contrast to the end plate potential the post-synaptic potential due to an impulse arriving at the synaptic knobs on a motoneurone belonging to a single afferent fibre is very small, the depolarisation being far below the threshold needed to excite an action potential in the motor axon ; but it lasts for some 10 milliseconds so that another impulse arriving within a few milliseconds can add a further post-synaptic potential to it, and so on. Thus the frequency of arrival of nerve impulses at the synapse determines the level of depolarisation of the motoneurone. A similar effect can be obtained at the muscle end plate by treatment with curare, using a dose which reduces the size of the end plate potential until with a single nerve impulse it is no longer large enough to excite the muscle fibre. With two or more impulses at intervals of a few milliseconds, the end plate potentials summate and reach a level sufficient to excite.

Another difference from a muscle fibre is that a motoneurone has synaptic connections with not one but a large number of afferent fibres. The degree of depolarisation of the motoneurone thus depends both on the number of afferent fibres sending impulses (spatial summation) and on the frequency of impulses in them (temporal summation). In experimental work where the afferent fibres are excited by electrical stimuli applied to a muscular nerve, spatial summation is demonstrated by varying the size of the volley and temporal summation by varying the interval between two volleys. Whenever in either of these ways the depolarisation of the motoneurone reaches a threshold value of about 30 mV. an action potential is excited. In more natural circumstances, when for example pulling on a muscle sets up a continuous asynchronous discharge of impulses in many afferent fibres, both types

of summation occur together and result in a more or less steady depolarisation of the motoneurone ; with a sufficient depolarisation a continuous discharge of motor impulses is set up in the axon. The frequency of discharge rises as the level of depolarisation increases. A similar discharge is observed if direct current is passed through a motoneurone in such a direction as to depolarise it. Myelinated nerve fibres do not in general respond to constant current by repetitive firing in this way ; they accommodate. Clearly there must be some part of the motoneurone or the axon near it which does not accommodate. The axon near certain sense organs is known also to behave in this way (p. 371).

Inhibition. Synaptic endings which depolarise the motoneurone and tend to excite it are not the only kind. There are also inhibitory synaptic endings on the motoneurone whose action is to cut short or oppose excitation. We shall see in the next chapter some of the circumstances which bring them into play, but this is a convenient place to deal with their mode of action. The effect of an inhibitory synapse is greatly to increase the permeability of the motoneurone membrane to potassium and chloride ions but not to sodium ions. As explained in Chapter 12 (p. 368), the effect of this is to hold the membrane at the resting potential and reduce the efficacy of currents tending to depolarise. Thus inhibitory synaptic action makes the motoneurone more difficult to excite, by reducing the effect of excitatory synaptic action.

When impulses reach the inhibitory synapses the membrane potential may or may not alter. If it is at the resting potential given by the potassium and chloride concentration ratios it will not move. If the motoneurone is partly depolarised, either by excitatory synaptic action or because it has begun to deteriorate, the microelectrode records a shift back towards the resting potential, *i.e.* a repolarisation. Thus inhibition by this mechanism is not to be thought of as the converse of excitation ; there is never a true hyperpolarisation. The chemical nature of the transmitter at inhibitory synapses in the motoneurone is not yet known. It has been shown, however, that both strychnine and tetanus toxin block inhibitory synaptic action on the motoneurone and this is suspected to be why they cause convulsions. Other convulsant drugs do not act in this way.

The type of inhibition that acts on motoneurones themselves is not the only kind of which we have evidence. There are other inhibitory fibres which appear to act by reducing the amount of transmitter released from the excitatory synaptic endings in motoneurones, possibly by reducing the size of the nerve impulse reaching the synaptic knobs. No doubt among the ways in which nerve cells excite and inhibit each other are many as yet unsuspected devices of this kind. Indeed, little is yet known about synaptic transmission in the central nervous system elsewhere than at the motoneurone. Substantial quantities of acetylcholine, adrenaline and noradrenaline (pp. 462–467), and the enzymes associated with these transmitters are found in the brain, and there are other indications that some brain synapses are cholinergic and some adrenergic. But as we have seen there are at least two unknown

chemical transmitters operating on the motoneurone alone, and no one knows how many await discovery elsewhere.

Electrical Synapses. It is worth remarking that we have no guarantee that all synapses do employ chemical transmitters. As we saw, the chemical mechanism at the end plate is a device to allow a small structure to excite a large one by amplifying the depolarising current. The same is obviously true of the excitatory synaptic knobs on a motoneurone. But if two structures are of comparable size there is no reason why one should not excite the other electrically if a low resistance path for depolarising current is provided where the two membranes come together. Such electrical synapses are known to occur in crayfish where the giant nerve fibre excites smaller motor nerves ; the same fibres have ordinary chemical synapses elsewhere. Nature may prove to have been equally opportunist in the mammalian central nervous system.

One-way Transmission. A characteristic of all vertebrate synapses yet investigated is that excitation only passes in one direction across them. (Two-way synapses of a highly specialised kind are found at the intersegmental junctions on the giant nerve fibres of the earthworm.) In chemical synapses the transmitter is only liberated by one side of the synapse and only acts on the other side. Electrical back-excitation does not occur because no low resistance path exists across the synapse. In the crayfish electrical synapse one-way conduction is ensured by an electrical rectifying action of the synaptic membranes ; current only flows easily in one direction. It is because of this one-way property of synapses that all nerve fibres, yet recorded from, normally carry impulses in one direction only. One-way synapses and one-way impulse traffic are only to be expected in the sensory nerves and tracts carrying sensory messages into and within the central nervous system, and in the motor tracts and motor nerves taking orders back to the muscles. These are the parts of the nervous system about which most is known at present. Once again there is no guarantee that elsewhere in the central nervous system all synapses will prove to be one-way or that all nerve fibres will be found to carry one-way traffic. So much remains to be discovered about the nervous system that generalisation beyond what is established by experiment is always unsafe.

CHAPTER 14

THE CENTRAL NERVOUS SYSTEM

Anatomy. Because the anatomy of the central nervous system is commonly less familiar, even to medical readers, than that of the rest of the body, all the structures subsequently mentioned in this chapter are briefly described in this section or are depicted in Figures 14.10, 14.11, 14.12 and 14.16. A more thorough knowledge can be acquired from the books listed in the Bibliography, but the non-medical reader who may be daunted by the large volume of detail in such works should remember that, although it is axiomatic that physiology must always rest on a firm basis of anatomy, so little of the physiology of the brain is understood at present that a relatively small amount of anatomical knowledge will be found adequate to support it. Outside the skull there are probably no organs or tissues in the human

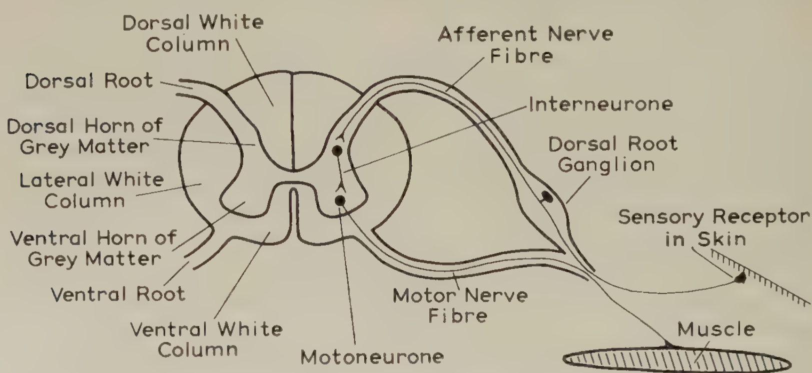


FIG. 14.1. Diagrammatic cross section of the spinal cord showing the H-shaped grey matter, the columns of white matter surrounding it, and the components of a simple reflex arc.

body, except perhaps the thymus gland, of which the function is not understood, at least in a general way, but in the brain the student must not allow himself to be nonplussed to find numerous large masses of nerve cells for which factual information about function is negligible. He may be warned, however, that speculation on these matters, masquerading as fact, in anatomical and physiological texts, is an old if little recognised branch of science fiction.

The central nervous system consists of the spinal cord, the part within the spinal column, and the brain, the part within the skull. The spinal cord retains the segmental structure of the primitive nerve cord, giving rise in each segment to a pair of spinal nerves, each formed from separate dorsal (sensory) and ventral (motor) roots (Fig. 14.1). Inside the cord is a core of nerve cells (grey matter) surrounded by the tracts of nerve fibres (white matter) that carry nerve impulses up and down the cord. The spinal cord is the seat of various reflexes, protective, muscular, visceral, sexual, etc., and of various automatic nervous mechanisms concerned in standing, walking, etc. It also transmits sensory information to the brain, and executes the

orders the brain sends it. In the higher vertebrates the spinal cord becomes increasingly subordinate to the brain, and the tracts of white matter running to and from the brain occupy a large part of it. The position of the principal spinal tracts is indicated in Figs. 14. 11, 14. 12, and 14. 16.

The vertebrate brain develops from three hollow swellings at the cranial end of the primitive neural tube, the forebrain, midbrain and hindbrain vesicles. The olfactory and optic nerves take origin from the front and back respectively of the forebrain ; all the other cranial nerves belong to the mid- and hindbrain (Fig. 14. 2). The main external features of the mammalian brain result from elaboration of this basic structure in two regions. (1) From the front of the forebrain two further hollow swellings develop, the cerebral hemispheres (Fig. 14. 2). The original cavity of the forebrain vesicle becomes known as the IIIrd ventricle and its extensions into the hemispheres are the two lateral ventricles. (2) In the roof of the front part of the hindbrain develops a large unpaired structure, the cerebellum.

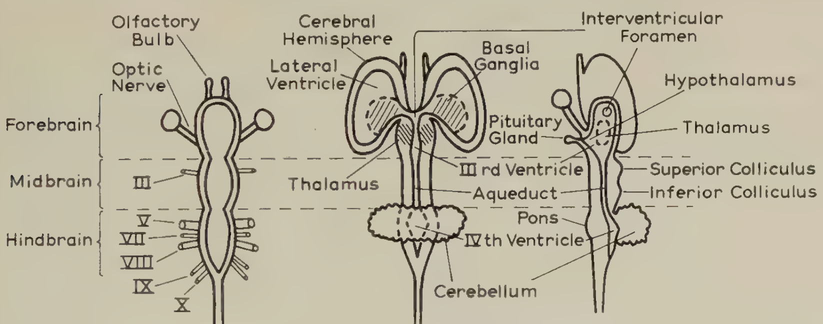


Fig. 14. 2. Idealised mammalian brain in two stages of development. Left : the three primitive brain vesicles in horizontal section and the principal cranial nerves. Centre and right : a later stage, in horizontal and in midline section (cranial nerves, except optic nerves, omitted).

Cranial nerves : III (oculomotor) motor to most of the eye muscles ; V and VII (trigeminal and facial) sensory and motor to face and mouth ; VIII (auditory) consists of cochlear and vestibular portions from the organs of hearing and balance respectively : IX and X (glossopharyngeal and vagus) motor and sensory to the alimentary and respiratory tracts and to the heart.

Tracts of nerve fibres running transversely to the cerebellum form a conspicuous bulge, the pons, on the ventral surface of the hindbrain. Behind the pons, where it tapers into the spinal cord, the hindbrain is known as the medulla oblongata, or medulla for short. The cavity of the hindbrain is named the IVth ventricle ; it is connected to the IIIrd ventricle by the aqueduct, the narrowed cavity of the midbrain vesicle. The mid- and hindbrain, excluding the cerebellum, are sometimes referred to jointly as the brainstem.

The cerebral hemispheres and the cerebellum are the parts of the brain that increase most in size as the evolutionary scale is ascended. In man they are so large that they hide most of the rest of the brain (Fig. 14. 3). The cerebral hemispheres and the cerebellum are the only places in the central nervous system where grey matter lies on the surface forming a " cortex." Elsewhere the grey matter is in the middle with the connecting tracts of white matter around it. Underneath the cerebral cortex, thick bundles of white matter connect the parts of the cortex to each other and to other parts of the brain. A massive tract, the corpus callosum, runs between the two cerebral hemispheres above the roof of the IIIrd ventricle (Fig. 14. 3). The main tracts connecting the cerebral cortex to the lower parts of the brain and to the spinal cord pass between the basal ganglia and the

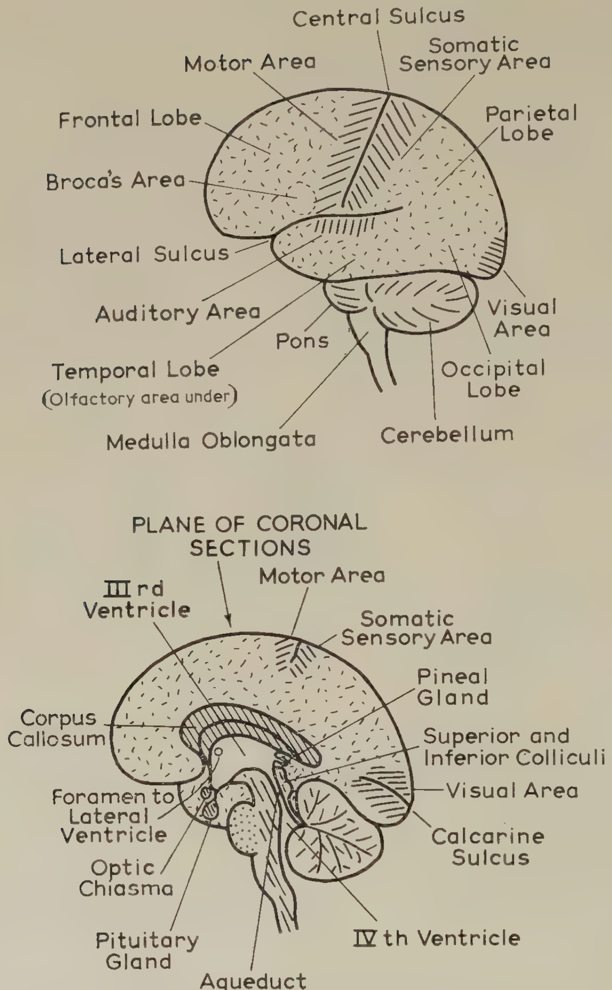


FIG. 14. 3. Lateral view and median sagital section of the human brain, with " projection areas " marked.

Above the pituitary gland and optic chiasma the lateral wall of the IIIrd ventricle is formed by the hypothalamus, above that by the thalamus. The arrow indicates the plane of the coronal (frontal) sections in Figs. 14. 11 and 14. 16.

thalamus (see below), where they are known as the internal capsule, and emerge on the ventral surface of the midbrain to form the cerebral peduncles.

The basal ganglia, or corpus striatum, is a mass of grey matter that develops in the floor of the lateral ventricle. It is the highest centre for motor functions, apart from the cerebral cortex. The thalamus is a large group of nuclei in the wall of the IIIrd ventricle. It is the highest centre for sensory activity apart from the cerebral cortex. Below the thalamus the floor of the IIIrd ventricle descends to meet Rathke's pouch growing up from the roof of the buccal cavity, to form with it the pituitary gland. The grey matter

in the walls of the IIIrd ventricle immediately above the pituitary constitutes the hypothalamus.

The brain stem contains a central core of grey matter, continuous with that in the spinal cord, from which arise the cranial nerves (Fig. 14. 2), but the segmental origin and the division of each nerve into motor and sensory roots seen in the spinal cord are not in evidence in the cranial nerves. Among the cranial nerve nuclei and the other specifically named structures in the brain stem there is also a large region where small groups of cells lie in a network of nerve fibres running in all directions, hence known as the reticular formation (Fig. 14. 16). This part of the brain is very fashionable among neurophysiologists, who do not always confine the term to the true reticular formation of the anatomists. Many functions are attributed to the reticular formation some of which are discussed in this chapter on pp. 437, 444 and 451. The respiratory and vasomotor centres (Chapters 4 and 2) are also claimed as part of the reticular formation.

Categories of Nervous Action

The central nervous system is the organ which directs the behaviour of animals, all those responses and activities by which an animal reacts to its environment and attempts to master it and flourish. In man and many other animals nearly all behavioural activities are carried out by the contractions of skeletal muscle. (This is not always so ; the skunk and the electric fish have other ways of getting the better of their enemies.) In addition to its outwardly directed activities the central nervous system has the job of running most of the internal systems of the body ; it controls, for instance, the endocrine, digestive, respiratory, blood circulatory and excretory systems. Most of these internal duties are discussed in the chapters dealing with those systems; in this chapter we are mainly concerned with the central nervous system as it determines the external behaviour of the animal through its muscles.

An animal's reactions to changes in its environment may differ greatly in complexity. Shining a light in the eye causes the pupil to constrict ; this is one of the simplest of nervous reactions, and historically one of the first that was clearly understood to involve a nervous pathway from the sensitive surface to the central nervous system and back to the effector muscle. At the other end of the scale the production of a Ph.D. thesis may equally clearly be a response to a change of environment. Those reactions that seem to be automatic and immediate, such as blinks, knee jerks or sneezes, are termed "reflexes." The other main category of muscular actions is those that involve conscious thought and decision called "voluntary movements." Typically a reflex is something the subject can't help doing, like coughing during a concert, or choking if someone touches the back of his throat with a feather, whereas voluntary actions are actions he could have refrained from if he had wished ; hence the satirical definition of a voluntary movement as a movement a patient makes when he is told to by a neurologist.

The line between reflex and voluntary is very indistinct. There are all levels of behaviour between the most voluntary and unpredict-

able, such as writing a Ph.D. thesis, and the most automatic, such as the pupillary reaction to light. Reading a book is obviously a voluntary activity, looking up when someone enters the room is practically automatic, blinking when a threatening movement is made towards the face, although it may be possible to prevent it by a strong effort of will, is so automatic that it is generally considered a reflex, blinking when an object touches the cornea is an indubitable reflex.

No one knows whether the nervous mechanisms of the most elaborate voluntary actions and of the simplest reflexes differ in principle or only in complexity and detail. No rational classification of the activities of the central nervous system is therefore possible in the present state of knowledge so whether we call some particular manifestation of nervous action voluntary or reflex, or by some other term, is purely a matter of descriptive convenience, and usages are bound to differ in so complex a subject.

An important category of semi-automatic reactions that are not usually called reflex are those that express emotion. Emotional movements of the face are certainly different from voluntary movements in their nervous mechanism because, as we shall see (p. 433), disease of the brain may paralyse voluntary movements of the face such as screwing up the eyes, baring the teeth, etc., and yet the patient may smile normally if amused.

So far we have spoken as if all behaviour were a reaction to some environmental stimulus direct or remote, but, of course, in ordinary life a good many of a man's or an animal's voluntary acts are likely to be in response to internal stimuli, thirst, hunger, and sexual and other instincts. Instances of simpler types of activity that appear to arise spontaneously, without any obvious external stimulus, are seen with blinking and with eye movements. Blinking may sometimes be reflex, as when something touches the cornea, or it may be a voluntary act. Blinking, however, goes on irregularly all the time while we are awake and for most of these occasions there is no apparent external stimulus and certainly no voluntary factor ; its function is presumably to prevent the cornea from becoming dirty or drying up. Such blinking certainly appears to be spontaneous and, indeed, cannot for long be suppressed by the will. This urgency is even more in evidence with the incessant shifting of the direction of gaze that goes on all the time we have our eyes open. It is much more difficult to avoid moving the eyes than it is to avoid blinking. The need to move the fixation point at frequent intervals is probably connected with the fading of contrast which is observed if the retinal image is not allowed to move (p. 562). At all events it cannot be because the world is full of interesting objects that provide a succession of irresistible reflex stimuli, for spontaneous eye movements if anything increase when looking at a blank surface ; indeed, in the dark it appears to be impossible to keep the eyes still.

Nothing is known of the nervous mechanisms responsible for such apparently spontaneous actions, and they will not be discussed further. They are mentioned in order to emphasise that the nervous system may

take a much more active part in shaping events than is implied in the simple idea of reflex action.

It will also be evident from the foregoing that there are nervous actions of which it is no use asking, is it a reflex?, expecting an answer yes or no. Thus, as we have seen, blinking may be reflex ; more usually it is apparently spontaneous, but it may be voluntary or even emotional in origin. It is best thought of as just a specific type of muscular activity, that may arise in a variety of ways.

These remarks serve to introduce the ubiquitous and important subject of posture. Posture is maintained by the same muscles that are used to carry out the ordinary outwardly-directed behavioural activities (reflex, voluntary, etc.) of the animal, but otherwise it is more akin to the internal homeostatic mechanisms such as breathing. Readers will be familiar with Claude Bernard's teaching that stable surroundings for the cells of the body as regards temperature, pH, ionic composition, etc., are prerequisite for a life of unrestricted activity, but a land-living mammal is just as disabled if it cannot stand up as if, owing to faulty breathing, the pH of its blood is wrong. Correct posture is not so much a behavioural activity itself as an initial state which is necessary before an animal can act effectively. As with breathing, etc., it is best not to ask whether posture is reflex or spontaneous but to regard it as a state of the body which is controlled by automatic mechanisms sensitive to the direction of gravity, the pressure of the ground on the feet, etc., in the same way that breathing is controlled by the concentration of carbon dioxide in the blood, etc. No one thinks of breathing as a reflex response to a smack on the buttocks at birth, nor is standing merely a response to the sound of the breakfast gong. Both activities so long outlast the stimulus, that how they are maintained has become of much more interest than how they were initiated. (Standing is not a reflex response to gravity, for gravity acts very constantly and since animals sometimes lie down and sometimes stand up it cannot be gravity that determines which they do. It is true that when for other, usually inscrutable, reasons an animal stands up the direction of gravity determines how it does it, but that is not the same thing.)

Locomotion (walking, running, etc.) is another complex action whose mechanism deserves to be studied in its own right, quite apart from the stimulus that sets it going. It is closely related to posture and subsequently the two will be dealt with together.

This completes the introductory discussion of the types of nervous activity that have to be taken account of in this chapter. Nothing has been said of such things as social behaviour or psychology for the very good reason that we are so far from coming to grips with the neurological mechanisms involved that they are not yet part of the subject matter of physiology.

Since Sherrington's famous, but for many people almost unreadable, book " The Integrative Action of the Nervous System " (1906, reprinted 1947) it has been conventional to emphasise integration or co-ordination as functions of the central nervous system, so much so that some authors

apparently feel that to describe a nervous centre as exercising an integrating or co-ordinating influence over some function or other absolves them from attempting any description of what it does. Sherrington himself, who was always prone to use words in highly specialised senses of his own, clearly used the word " co-ordination " with a very extended meaning, as the following sentences, with which he opens an article on the brain, show. " The nervous system has as its function the co-ordinating of the activities of the organs one with another. It puts the organs into such mutual relation that the animal reacts as a whole with speed, accuracy and self-advantage, in response to the environmental agencies which stimulate it." This quotation can be regarded as defining " co-ordination." For Sherrington, it is apparently that activity which is the function of the nervous system. Other passages show that he used " integration " in much the same sense. To avoid misunderstanding, the words " integration " and " co-ordination " are not used in this chapter.

Reflex Action

The word " reflex " implied originally that nervous messages travelling up sensory nerves were " reflected " in the central nervous system and passed out again into motor nerves. Ideas of this kind were current in the seventeenth century ; an instance of reflection given by Descartes is the involuntary blink a man gives if a threatening gesture is made towards his face. Some of the first clear-cut experiments were made by Stephen Hales, the rector of Teddington, Middlesex, in about 1730. He used a spinal frog, *i.e.* a frog that has been decapitated so that the spinal cord is the only part of the central nervous system remaining. When such a preparation is suspended the legs hang down limply, but if one foot is pinched that leg is drawn up as if to remove the foot from the injurious stimulus. This withdrawal reaction (the flexion reflex) is lost if the spinal cord is destroyed, a result which strongly suggests that the nervous pathway from the skin of the foot to the flexor muscles runs via the spinal cord.

Hales' experiment does not definitely exclude other more far-fetched possibilities, such as that the site of reflection is really in the lumbar nerve plexus but that the nerve plexus is under the influence of the spinal cord and ceases to reflect if the cord is destroyed. (Something not dissimilar is now known to occur in the axon reflex from skin on to cutaneous blood vessels, see p. 68). But such alternatives were made superfluous by the celebrated experiments of Magendie and Bell on the functions of the spinal nerve roots, at the beginning of the 19th century. They showed that, if a ventral root was cut through, excitation of the peripheral stump caused widespread muscular contraction, but excitation of the central stump connected to the spinal cord was without obvious effect ; in particular the animal exhibited no signs of sensation. The peripheral stump of a divided dorsal root gave no muscular movements (and, of course, no signs of sensation), but the central stump caused obvious pain and elicited muscular movements (which disappeared if the ventral roots were divided too). These experiments showed that the ventral roots were exclusively motor, with no sensory function, and that the dorsal roots were exclusively

sensory but could excite movements by " reflection " in the spinal cord.

Before Magendie and Bell it was not clear that motor and sensory nerves were distinct. The real significance of their experiments was to show that the nervous system uses separate channels for input and output—the most important single principle in the organisation of the nervous system. Nowadays the idea of motor and sensory fibres is so familiar that it is easy to forget that Magendie and Bell discovered their separate existence and did not merely demonstrate an anatomical fact about their mode of origin from the spinal cord. As regards reflex action, the fact that motor and sensory fibres maintained a complete separation of function until they reached the spinal cord showed that the connection between them, the site of reflection, must be in the cord.

The anatomical parts of the simplest idealised reflex pathway (Fig. 14. 1) are : a sensory receptor connected to a sensory (afferent) nerve fibre running to the central nervous system, a motor (efferent) nerve fibre running from the central nervous system to a muscle, and a synaptic connection in the central nervous system between the sensory fibre and the motoneurone. This synaptic connection may either be direct or, more usually, via one or more additional nerve cells called interneurones. In all real reflex pathways, of course, many receptors and many sensory and motor fibres are involved.

The central connections between input and output are seldom simple enough to give the impression that sensory impulses are merely " reflected " into motor channels. An instance that might suggest reflection is the constriction of the pupil when light is shone into the eye. With the majority of reflexes, however, the pattern of discharge of motor impulses differs, often very widely indeed, from the pattern in the sensory nerves that excite them. A relatively simple reflex is the corneal reflex : a touch on the cornea, *e.g.* with a wisp of cotton wool, causes a blink. The sensory impulses for this reflex travel in the ophthalmic division of the trigeminal (Vth cranial) nerve to the medulla (see Fig. 14. 2), and motor impulses leave from nearby in the facial (VIIth cranial) nerve which supplies the orbicularis oculi—the muscle responsible for blinking. There is clearly a formal sense in which impulses may be said to find their way from the cornea via the facial nerve nucleus back to the orbicularis, but such a mode of description certainly tends to conceal the significant fact that a blink is an entity, almost an all-or-none event, a stereotyped pattern of motor discharge which differs little whatever type of stimulus is used to evoke it. Hence it is more illuminating to say that touching the cornea triggers off a blink, rather than that it reflexly excites the orbicularis. As Hughlings Jackson (1835–1911), the greatest of all neurological thinkers, put it, even in the simplest reflexes the nervous system " thinks " in terms of movements, not muscles—the corneal reflex results in a blink, not just in some contraction of the orbicularis. Nothing is known of the details of the motor mechanism that organises a blink.

Several examples of more complicated reflexes are found at the

head end of the digestive and respiratory tracts, swallowing, vomiting, sneezing, coughing, etc. In swallowing the problem is to get a mouthful of food or drink rapidly into the œsophagus without any entering the nose or the larynx. The elaborate sequence of muscular operations by which this is achieved has been described in Chapter 5 (p. 149). When voluntary activity is in abeyance, *e.g.* in stuporose human beings or in animals from whom the cerebral hemispheres have been removed (decerebrate preparations), reflex swallowing can be induced by placing small pieces of food on the back of the tongue. Vomiting, sneezing, etc., are also self-contained sequences of muscular acts which are triggered by sensory stimuli.

A different type of complexity is seen with micturition in the spinal dog. In a dog in which the spinal cord has been functionally separated from the brain by transection in, say, the thoracic region, the bladder, after some days, empties itself reflexly when it is full (see also p. 275). Emptying of the bladder is itself another example of a definite act triggered by sensory impulses rather than a reflection of these impulses to the effector organ, but what is of particular interest is that the spinal dog, when it has finished passing its water, may wag its tail, just like any ordinary dog. Thus this elaborate gesture involving many muscles, none of them anything to do with the bladder, is clearly a part of the reflex act of micturition.

The examples so far discussed make it evident that on the motor side the reflex centres of the central nervous system deal in terms of complete muscular acts. It is natural to ask if there is any corresponding generalisation to be made about the sensory side of reflex action. Take the case of the reflex blink in response to a threat. If we speak of this reflex as a matter of impulses finding a pathway from retina to orbicularis muscle we shall be on safe ground, but this formulation disregards the fact that only a spatiotemporal pattern of impulses in many thousands of optic nerve fibres which the brain recognises as a threatening object rapidly approaching the face will evoke a blink. This is known to be a task that requires the highest centres for sensory interpretation in the cerebral cortex, for the blink to a threat from one side may be abolished by a small lesion of the opposite parietal lobe (see Fig. 14. 3, and p. 423) which causes no detectable disturbance of vision. Evidently in this reflex by far the most difficult part is the process of recognising the threat ; this is presumably why the delay after a threat is made before the blink occurs (variously known as the latent period or the reaction time) is of the order of 250 milliseconds, whereas the blink to a touch on the cornea takes only some 40 milliseconds. For the purposes of this reflex, therefore, impulses in the optic nerve are of no importance as such, but only become important when their message has been decoded into whatever pattern of nervous processes it is that constitutes the nervous representation of a threat ; the threat is the significant thing, not the impulses. To adapt Hughlings Jackson, the central nervous system " thinks " in terms of the causative physical events not in terms of sensory receptors.

The blink reflex is an extreme example of this principle because the process of recognition is so complex and the motor mechanism relatively so simple, but at lower levels in the central nervous system sensory recognition is still important, chiefly in reflexes initiated from the skin. Elsewhere it is not so in evidence ; *e.g.* when impulses from distension receptors in the bladder arrive at the spinal cord presumably they can only mean one thing ; no problem of recognition arises. But the skin is played upon by many diverse influences, and the appropriate reflex responses can only be made if the central nervous system can distinguish the various stimuli.

In a spinal dog or cat two opposite reflexes can be elicited from the pads of a foot. Pressure by a flat object on the pads causes a brief powerful extension of hip, knee and ankle, the extensor thrust reflex (tentatively identified by Sherrington as an element in galloping), whereas pinching a pad (or any other form of painful stimulation) gives the withdrawal or flexion reflex of the limb, already mentioned in the frog. The stimulus for the extensor thrust presumably excites touch and pressure receptors. A pinch which elicits flexion excites pain fibres as well as touch and pressure. Experiment shows, however, that any noxious stimulus which is likely to cause pain will elicit the flexion reflex, so the essential thing about a pinch must be the pain, and this must be easy enough to recognise because it is carried by special pain fibres. Protective reflexes caused by noxious stimuli always seem to get priority in the nervous system (for obvious practical reasons) whence it appears that the recognition of pain carries with it an automatic refusal to recognise other stimuli (in this case touch and pressure) or in some other way overrule them.

The recognition of the stimulus that calls for an extensor thrust is not so straightforward a matter, however. It is not merely a matter of impulses in touch and pressure fibres in the absence of pain. Sherrington never succeeded in eliciting an extensor thrust reaction by electrical stimulation of the sensory nerves supplying the pads although there is no reason to doubt that he was exciting touch and pressure fibres. (Pain fibres, being of small diameter, require a larger electrical stimulus [see p. 374] so it is not difficult to avoid exciting them.) It seems likely that recognition of pressure on the pads depends on a distinctive pattern of discharge from touch and pressure receptors in different positions on the pads, and also perhaps on a distinctive distribution in time of the impulses from each receptor, and that electrical stimulation fails to imitate this spatio-temporal pattern of discharge.

At all events this difficulty in eliciting reflexes by electrical excitation of nerve trunks is a widespread and interesting phenomenon that has attracted attention since the middle of the nineteenth century. With large stimuli reflexes due to excitation of pain fibres are obtainable, but it is probably true to say that no other reflexes can be elicited by electrical stimulation of skin nerves (muscle nerves are different, as is discussed on p. 413). The explanation offered is that for most skin stimuli recognition depends on characteristic patterns of discharge

in several nerve fibres which cannot be reproduced by electrical stimuli to nerve trunks; whereas pain stimuli are recognised because they travel in pain fibres and the pattern of discharge is unimportant; hence shocks large enough to excite pain fibres successfully excite pain reflexes.

Subjective observations by human subjects lend plausibility to this interpretation. Repetitive electrical stimulation of cutaneous nerve trunks is easily performed with an induction coil (Fig. 12. 3) through pad electrodes applied to the skin over the nerve. With shocks that are not so large as to cause pain all that is felt is a vague tingling and sensation of tightness over the area of skin distribution of the nerve, not unlike "pins and needles." The feelings aroused never correspond to any sensation that could result from natural stimulation of the skin surface. This failure to evoke recognisable sensations cannot be due to any inadequacy in the total number of impulses but must be because of their unnatural arrangement. Apparently the reflex centres of the spinal cord likewise cannot make sense of the synchronous volleys of impulses that result from electrical stimulation of skin nerves and they therefore give no reflex responses. With stronger shocks the situation alters; in the human subject pain sensations are aroused which are recognised as similar to those caused by ordinary painful stimuli; likewise in animals ordinary pain reflexes are elicited.

The general conclusions from the foregoing are that the reflex centres only recognise patterns of afferent impulses which correspond to naturally occurring physical stimuli, and that these stimuli when recognised only elicit definite movements each of which is an appropriate response to its stimulus. The reflex centres apparently have no truck with nonsense either on the sensory or the motor side; nonsensical afferent volleys are disregarded and nonsensical contractions of muscles are not produced. A corollary of this last statement is that it ought to be impossible to elicit two incompatible reflex movements simultaneously, for the resultant movement would be nonsensical in the sense that it would be inappropriate to either stimulus. This does appear to be the case; it is said to be impossible, for instance, by any kind of stimulation of the foot to elicit the extensor thrust and the flexion reflex simultaneously and so obtain an ineffectual simultaneous contraction of both the flexor and extensor muscles of the limb. Nothing definite is known of the details of the connections between nerve cells that are responsible for the properties summarised in this paragraph.

Spinal Reflexes. The central nervous system is so immensely complex that to analyse how it works is bound to be extremely difficult. On the intact animal it is not possible to proceed by the classical scientific method of varying only one factor at a time while all others are held constant, because an animal's response to a particular change in its environment often depends on things that happened to it long before the experiment started. Even its reflex responses tend to be variable and unpredictable, and may furthermore be disturbed by apparently spontaneous activity. These difficulties can be minimised by confining attention to the reflex responses of the spinal cord after

it has been severed from the brain to exclude voluntary and spontaneous interference.

For a time after spinal transection it is impossible to elicit any, or almost any, reflexes from the isolated part of the spinal cord below the section. This is called spinal shock. After shock passes off the reflexes tend to become brisker and more easily provoked than they were originally. The higher an animal is in the evolutionary scale the more profound the shock, the longer it lasts, and the smaller the number of reflexes that eventually reappear, although ultimately some of these reflexes may be greatly enhanced. The decapitate frog recovers from spinal shock in a minute or two. In the cat or dog some reflexes reappear in a matter of minutes ; most of those that are going to appear at all do so in a few days, but the ease with which they can be elicited may go on increasing for months after. In the spinal monkey and in cases of accidental division of the spinal cord in man (spinal man) the return of responsiveness does not begin for several days and progresses very slowly.

The nature of spinal shock is still unknown ; it is clear that shock is a bad name for something that may last for years and that grades imperceptibly into enhancement. It is not the mechanical damage that causes shock for, in an animal that has recovered from shock, a second section immediately behind the first causes no return of shock. Sherrington showed in the monkey that at a time when the lower limbs were entirely unresponsive to massive stimuli (such as repetitive electrical stimulation of large nerve trunks or extensive burning of the sole of the foot) which would normally cause violent limb reflexes, movements could still be obtained with ease by exciting electrically the tracts of the spinal cord exposed at the cut surface. This suggests that the spinal neurones are not so much inexcitable as deprived of excitation normally reaching them from higher levels in the nervous system. (This would explain why there is no trace of shock in any part of the nervous system in front of the section). The more advanced the animal the more its behaviour is dictated by its brain and the larger the deficit of excitation in the spinal cord after it is cut off from cerebral influences. During recovery certain reflex pathways, in some unknown way, increase their power of exciting and overcome the deficit. The larger the deficit the longer it takes to fill.

The reflexes that recover seem only to be those that in the ordinary life of the animal are little or not at all under voluntary control. In man almost every action is voluntary or requires voluntary permission, and a spinal man is left, below the level of the lesion, with little more than flexion reflexes (p. 404), tendon jerks (p. 411) and a number of visceral (sympathetic and parasympathetic) responses, notably automatic emptying of the bladder and rectum, sexual reflexes (erection and seminal emission) and vasomotor responses to pain and to a full bladder (see p. 460). Many of these become unnaturally brisk and easily provoked. In a proportion of cases the stretch reflex (p. 407) returns and after some years brief periods of spinal standing may be possible. In a

spinal dog or cat, apart from similar visceral responses there are a large number of protective reflexes (of which the best known are the flexion reflex and the scratch reflex (p. 405)), some rather half-hearted attempts at standing and walking, several fragmentary reflexes of less easily recognisable purpose (such as the extensor thrust (p. 399) and the stretch reflex (p. 407)) which appear to be elements of postural or locomotor mechanisms or of the nervous mechanisms that control muscular contraction, and little else.

It is a price that the experimenter has to pay for getting all the variables of the experiment under his control that the isolated spinal cord of the ordinary laboratory mammals does not have a very large or interesting repertory of reflexes, but to offset this the simplicity of spinal reflexes has been turned to great advantage for purposes of analysis, and most of what is known of the detailed behaviour of nervous pathways has come from the study of a few spinal reflexes. Sherrington, who pioneered in this field in the last years of the nineteenth century, was impelled by the belief that reflexes were the units of nervous action and that more complicated types of behaviour would prove to be compounded of simple reflexes. Although what he and his pupils discovered about individual reflexes and the way reflexes interact (for instance the way in which the protective flexion reflex overrides, or inhibits, other reflexes) is of very great interest and importance, the attempt to show that complex acts like walking are made up of a succession of simpler reflexes was not equally successful and may (as is further discussed on pp. 407 and 417) prove to be mistaken. Hence the reader must not be disappointed to find that a knowledge of what goes on in the simpler spinal reflexes, of the nature of synaptic action and so forth, does not throw much light on the mechanism of more complex reflex actions, let alone on the mode of functioning of the higher parts of the brain.

The spinal reflexes which have been investigated in greatest detail and from which most generalisations about reflexes are conventionally drawn are the flexion reflex, the scratch reflex, and the stretch reflex.

The Flexion Reflex. The flexion or withdrawal reflex of the limb to a painful stimulus has already been referred to more than once ; some of its characteristics as seen in the cat will now be described more fully. The reflex exhibits a threshold ; to elicit it any stimulus must reach a certain intensity. The interval between the time of application of the stimulus and the start of the reflex contraction (the latency of the reflex) is of the order of 30 milliseconds, much longer and more variable than the latency of contraction when a stimulus is applied to a motor nerve. A stimulus is more effective the larger the area it covers and the longer it lasts (spatial and temporal summation). As the stimulus is increased in intensity the flexion movement becomes more vigorous and more extensive ; thus a weak stimulus may cause movement of toes and ankle only, while with a very strong stimulus the knee and hip flex as well (irradiation). With strong stimuli the

limb may remain flexed for a second or more after the stimulus ends (after-discharge). The words in brackets are the special terms introduced by Sherrington.

In the most clear-cut demonstrations of the above properties Sherrington and his colleagues used electrical stimulation of peripheral nerves to elicit the reflex and recorded the contractions of an individual muscle (tibialis anterior) by detaching its tendon from its insertion and connecting it instead to an isometric lever (p. 480). The reflex latency could be measured easily with this arrangement. After making allowance for conduction time to and from the spinal cord the delay in the cord itself was found to be a minimum of about 5 milliseconds (at times it is very much longer). Spatial summation was shown by applying shocks to two nerves, both shocks just too weak by themselves to elicit a reflex (subliminal), but successful if applied together. When both shocks were large enough to produce small reflex contractions, applying them together gave a tension greater than the sum of the tension produced by stimulating the two nerves separately (spatial summation again). When, however, both shocks were increased to give the largest possible reflex contractions, applying both together often gave a tension little greater than with either separately (occlusion), because the muscle was already giving a nearly maximal contraction with the single stimuli. Temporal summation was shown by applying two subliminal shocks in rapid succession to the same nerve ; when the interval between them was less than some 20 milliseconds a reflex contraction resulted.

The fully developed flexion reflex is accompanied by extension of the opposite leg, the knee and ankle straighten out and the limb is thrust backwards at the hip. This *crossed extensor* reflex has a much longer latency (40–100 milliseconds) than the flexion reflex proper, and once it starts it takes a second or two to build up to its greatest strength (recruitment), in contrast to the rapid onset of flexion. Crossed extension still occurs when the dorsal roots of that limb have been cut. The crossed extensor reflex probably represents the first part of a step away from the noxious stimulus.

Temporal summation, as described above for the flexion reflex, implies that the first shock causes a change in excitability somewhere on the pathway of the reflex which persists for several milliseconds to summate with the effect of the second shock. Sherrington called this persisting change " central excitatory state." It seems to have been taken for granted that the alterations of central excitatory state during the flexor reflex took place in the motoneurones of the flexor muscles, in the motoneurone pool as it was termed. Electrical recording with fine electrodes inside the spinal cord has now shown that the afferent nerve fibres that elicit the flexion reflex (and in fact all skin afferents), end on interneurones in the dorsal horn of grey matter (Fig. 14. 1) and do not themselves run through to the motoneurones in the ventral horn. Afferent impulses from skin arriving in the dorsal horn start up potential changes there lasting for up to a tenth of a second. These

slow potentials suggest that there will be plenty of opportunity for temporal summation in the dorsal horn.

A reflex contraction of tibialis anterior can be produced by exciting any one of a number of peripheral nerves in that leg. Sherrington pictured impulses from these various sources "converging" on to tibialis motoneurones, which formed a "final common path." (As we have just seen the site of convergence is actually the interneurones of the dorsal horn.) The concept of the final common path found wide application in his writings, but nowadays, although true in a certain formal sense, it is a less illuminating doctrine than it previously appeared. For one thing, by implying that in reflex action afferent impulses find their way through the spinal cord eventually to emerge via the motor roots, it speaks the language of simple reflection and therefore tends to distract attention from the view preferred here (and endorsed incidentally by Sherrington himself) that the spinal cord "thinks" in terms of organised movements; other objections will appear later (pp. 443–4).

Withdrawal Reflexes in Man. Withdrawal reflexes can be elicited from the leg of a healthy subject by painful stimuli applied to the skin. Kugelberg and his co-workers have shown that the precise kind of movement obtained varies with the site of stimulation. As a convenient and reproducible stimulus they used a burst of high voltage electric shocks delivered through a pair of electrodes pressed against the skin. With the stimulus applied to the underside of the big toe the response obtained was flexion at all joints: upward movement of the toes, flexion of the ankle, knee and hip and forward flexion of the trunk, as illustrated in Fig. 14. 4. When the stimulus was moved further back on to the sole of the foot the toes moved downwards instead of upwards, the other joints flexing as before. The same response can be obtained by firmly stroking the sole. With the electrical stimulus under the heel the toes again moved downward but now the ankle extended too, the other joints again flexing. With the stimulus on the buttock there was extension of the trunk and hip, some flexion of the knee, extension of the ankle, and downward movement of the toes, as shown in Fig. 14. 4.

The effect of these movements in each case is to withdraw the stimulated point from the stimulus; but in addition they contain components which are clearly not part of the withdrawal, downward movement of the toes and extension of various joints. In a standing subject, the effect of these components of the reflex would be to press the toes downwards and assist to maintain standing. This only does not occur when the stimulus is applied to a toe, when it would be inconsistent with effective withdrawal. Hence it is thought that the reflex responses of the human leg to painful stimuli have the dual function of defence with maintenance of posture.

When the corticospinal motor tracts (p. 431) from the brain (without which a man cannot stand) are damaged, the components of the withdrawal response which appear to assist standing disappear. Electrical stimuli to the sole of the foot in such patients gave exactly the same

response (*i.e.* upward movement of the toes and flexion at all other joints) as stimulation of the underside of the toe in healthy subjects. Firm stroking of the sole of the foot in such patients also causes an upward movement of the toes. The same response is found in infants before they begin to stand up, and before the corticospinal tracts receive their myelin sheaths.

In spinal man an upgoing big toe accompanied by flexion of the leg is one of the first reflexes to appear during recovery from spinal shock. Over a period of months the threshold of the reflex falls and flexion

FIG. 14. 4. Withdrawal responses to painful stimuli applied to the underside of the big toe and to the buttock. The filled-in areas mark the trunk muscles that contract. (Kugelberg, Eklund and Grimby, *Brain*, vol. 83, p. 394, 1960.)

becomes more vigorous, until eventually violent flexion may result from trifling stimuli anywhere on the limb.

The response of the big toe to a firm stroke on the sole of the foot with a blunt point (the plantar response) is of great importance in clinical diagnosis, as an upgoing big toe is often the first unequivocal evidence of disease of the motor pathways (pp. 431, 434). The sign was first described by Babinski in 1896. An upgoing big toe is part of the general flexion reflex of the limb, but, although morphologically a movement of flexion, in anatomical nomenclature it is extension. Hence an upgoing big toe is referred to by neurologists as an extensor plantar response. A healthy plantar response is said to be flexor.

The Scratch Reflex. Some days or weeks after spinal section in the cervical region, a dog will respond to irritation of the skin over its

shoulder by scratching at the place with its hind foot. The movements made (Fig. 14. 5) closely resemble the familiar scratching of a normal dog, consisting of a series of rhythmical strokes of the whole leg at a frequency of about five per second. A good form of stimulus is a series of light touches with a pointed object. A single touch never succeeds; several dozen (at a rate of a few per second) may be necessary if they are all applied at one spot, but the total number diminishes if two or more spots are touched simultaneously; the most effective method is to touch a row of spots consecutively, thus imitating the pattern of

Fig. 14. 5. The scratch reflex in the spinal dog. The amount of the hip flexion recorded on a smoked drum by means of a string connected to the writing lever. The reflex was elicited by electrical stimulation of a point on the skin on the animal's back. The period of stimulation is shown by the descent of the signal line beneath the time scale. (Sherrington, " The Integrative Action of the Nervous System ": New York 1906; Cambridge 1947.)

stimulation normally caused by a moving flea. Once scratching has begun it continues for some seconds after the stimulus is withdrawn. (Readers are invited to apply to the above properties the Sherringtonian terminology given with the flexion reflex.)

The scratch reflex cannot be elicited by electrical excitation of cutaneous nerve trunks or of dorsal roots; apparently such volleys are not recognised as objects to be scratched at; but small shocks applied through a fine pin (electric flea) pushed just into the epidermis are successful. Not very surprisingly perhaps the frequency of scratching movements proves to be independent of the frequency of the electric

flea bites that cause them. This is one of the usually quoted formal differences between (repetitive) movements elicited reflexly and by direct excitation of a motor nerve. Another is fatigability: if the scratch reflex is elicited several times from the same point of skin it becomes weaker and more difficult to obtain. It may be useful to the animal to disregard a source of irritation it does not succeed in removing, so possibly the phenomenon is more akin to adaptation than to what is ordinarily thought of as fatigue. At all events the "fatigue" is not of the motor part of the reflex mechanism, for if the stimulus is moved to another point on the skin the reflex immediately returns in full strength. (It is not in general true that reflexes are readily fatigable. The cough reflex may persist with undiminished sensitivity after it has wholly exhausted the subject. The stretch reflex (p. 407) is also believed to be less fatigable than the muscles it employs.) If, during a scratch reflex, a painful stimulus is given to the leg, scratching immediately ceases and is replaced by a flexion reflex. This is an example of a protective reflex taking precedence by actively inhibiting a less urgent reflex.

The mechanism which generates the rhythm of scratching is not known. The frequency does not depend on the strength of the stimulus, or on the amplitude of the beat. Sherrington showed that if the amplitude is reduced by a carefully graded painful stimulus to a leg nerve, not quite strong enough to inhibit scratching completely, the frequency is unaltered. He also found that cutting the dorsal roots of the segments supplying the limb reduced the accuracy with which the scratching was directed to the point irritated, but did not alter the frequency. These experiments suggest that the rhythm is determined by a pacemaker in the spinal cord rather than by a sequence of flexor and extensor reflexes set up alternately by the moving limb itself; we shall see later (p. 417) that there is similar evidence for the slower rhythm of stepping.

The Stretch Reflex and the Tendon Jerk. So far we have dealt only with reflexes of obvious functional significance, elicited from the skin, the nose and throat, the cornea, the bladder, etc. In all cases the stimulus for the reflex would also have caused a conscious sensation in a normal human subject, pain, touch, tickle in the throat, etc. We now come to a very different reflex found in muscles. Muscles, like skin, are supplied with numerous sensory nerve fibres connected to sensory receptors, but unlike skin the impulses set up by these receptors do not give rise to conscious sensations. The knowledge we possess of the position of our limbs in space, position sense, comes from sense-endings in and around the joints, not from muscle.

Strictly, receptors and nerve fibres that do not cause sensations should not be called sensory, but this usage is accepted along with "sensory root" for dorsal root, etc., and causes no trouble if care is taken to make the intended meaning clear when ambiguity might arise. Other familiar examples of sense-endings that do not affect consciousness are the blood-pressure receptors in the carotid sinus (p. 45) and the vagal receptors signalling inflation of the lungs (p. 126).

Sense endings in muscles and those associated with joints were classed together as "proprioceptors" by Sherrington. The term is not used here as it tends to obscure the distinction that ought to be drawn between their functions.

The receptors in muscles elicit reflexes but, again unlike the reflexes from the skin, they are not of unequivocal function. The most important and best understood muscular reflex is the stretch reflex : when a muscle is stretched by pulling on it a reflex contraction is set up which opposes the pull. The stretch reflex can be demonstrated on intact animals and human beings ; it only reappears in the spinal dog after some weeks and in spinal man after a year or two. For analytical purposes it is usually studied on a decerebrate preparation, *i.e.* an animal whose cerebral hemispheres have been removed by making a hole in the skull under anæsthesia, cutting through the brain stem, usually just below the superior colliculi (Figs. 14. 3 and 14. 16), and scooping out the whole forebrain. Such a preparation has the advantage that the stretch reflex is present in the extensor muscles at once ; indeed it is exaggerated, leading to a state known as *decerebrate rigidity*, which is mentioned again on pp. 416, 444. Decerebrate animals are also convenient because without a forebrain they cannot (it is generally accepted) feel pain and so no anæsthetic is required ; also as the medulla is intact they breathe and maintain their blood pressure. The chief attention they require is to be kept at 38° C., because they have lost the automatic temperature control centres in the hypothalamus (p. 590).

In a decerebrate cat the stretch reflex is chiefly present in the extensors, the muscles that straighten the limbs. When such a muscle, *e.g.* the soleus, one of the ankle extensors, is cut away from its insertion and connected via its tendon to a device for measuring tension it is found that any attempt to elongate the muscle results in a rapid rise in the tension it develops, and at the same time electrodes on or in the muscle record numerous action potentials (Fig. 14. 6). This active contraction is reflex, because it disappears if either the motor nerve to soleus or the appropriate dorsal or ventral roots are cut through. Stretching the muscle after cutting its nerve then only causes a much smaller rise in tension due to the simple passive elastic properties which muscle, like any other piece of soft tissue, possesses. It is clearly established by other experiments that it is the change in length of the muscle that excites the stretch reflex and not the change in tension. Except when reflex excitability is artificially raised the stretch reflex is private to the muscle stretched, no other muscles contract ; and it persists in that muscle when the whole skin of the limb and every other muscle in the limb is denervated by systematic section of the nerves.

In human patients the neurologist detects the absence, presence or exaggeration of the stretch reflexes by assessing the "sense of passive resistance" he obtains when he moves the joints of the limb about while the patient relaxes and initiates no voluntary movements of his own. To express the results the somewhat unfortunate word "tone" is used. The normal degree of passive resistance is called normal

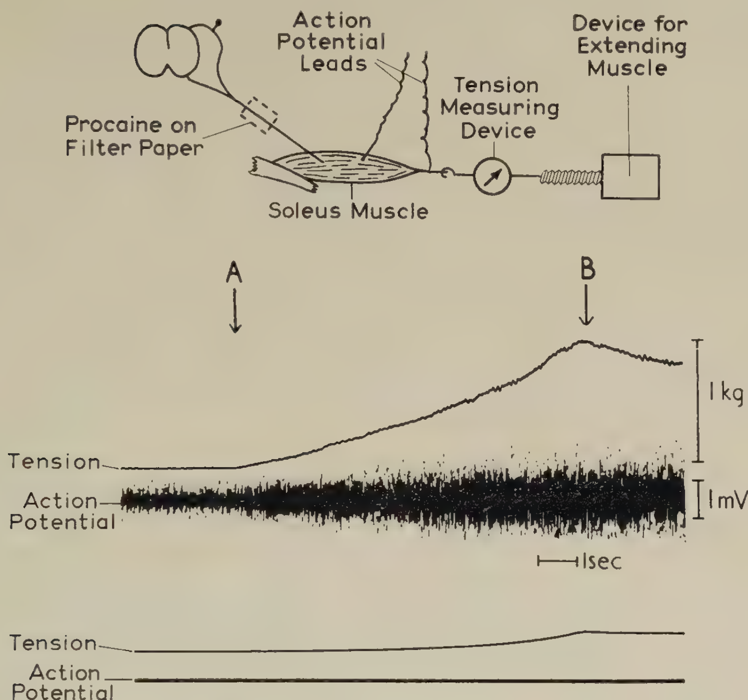


FIG. 14. 6. The stretch reflex in the soleus muscle of a decerebrate cat. The diagram above shows the experimental arrangements. The bones of the leg are clamped to a stand and the severed tendon of the soleus muscle connected to a device for recording muscle tension. To stretch the muscle the whole tension recording device is moved by means of a motor driven screw. The electrical response of the muscle is led off by electrodes on the muscle. (The application of procaine refers to Fig. 14. 9.)

The records below show the rise in tension that occurs when the muscle is extended 13 mm. at a rate of 1·7 mm. per second. Extension starts at A and stops at B. At the same time a great increase in the electrical activity in the muscle occurs, showing that the increase in tension is associated with an active reflex contraction of the muscle. The lower records, taken during a similar extension after the stretch reflex had been abolished, show the absence of electrical response and the much smaller rise in tension caused by the passive elastic properties of the muscle. (P. B. C. Matthews, unpublished records.)

tone ; diminished sense of passive resistance is diminished tone, the muscles are hypotonic or flaccid ; increased sense of passive resistance is increased tone with hypertonic or spastic muscles.

Opinions about the function of the stretch reflex are undergoing revision. It was once supposed that the stretch reflex was a characteristic of postural antigravity muscles, but it is now clear that the stretch reflex is active during reflex and voluntary as well as during postural contractions, and occurs in flexor as well as in extensor muscles. The view adopted here is that the stretch reflex is part of the general

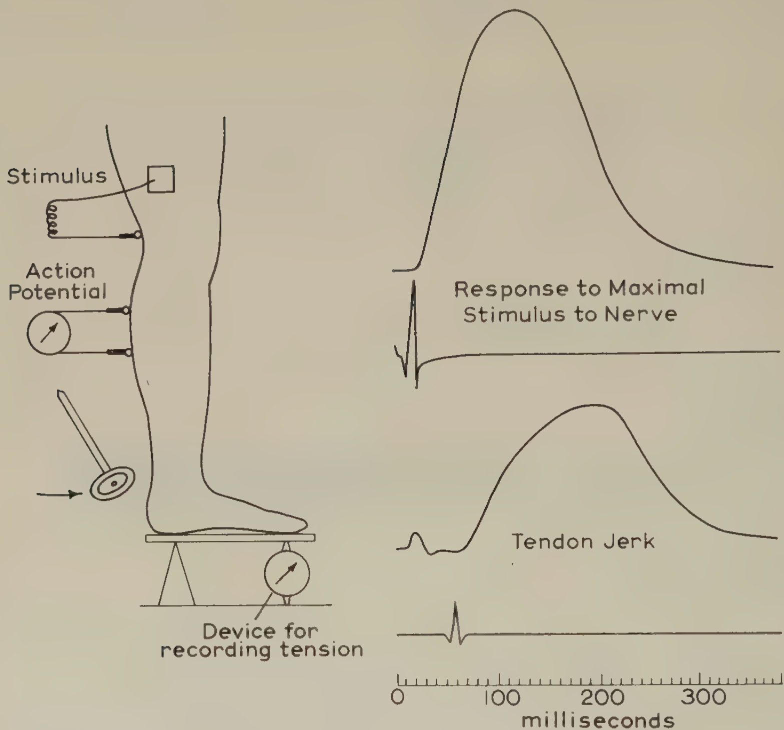


FIG. 14. 7. The human ankle jerk. The reflex, which is obtainable in almost all healthy persons, is elicited by striking the tendon of the ankle extensors (the soleus and gastrocnemius muscles) a sharp blow with a rubber-covered hammer ; it consists of a twitch contraction of these muscles. To record the contraction the foot rests on a board hinged under the heel ; contraction of the ankle extensors causes an increase in the downward pressure exerted by the front of the foot on the board, which is recorded on a cathode ray oscilloscope. The action potential of the extensors is led off by surface electrodes on the calf and a record of it appears below the tension record.

The oscilloscope time base is triggered by contact of the hammer with the skin over the tendon. The blow itself causes a brief rise in tension which appears as a small hump before the reflex twitch begins. The latency of the reflex in this subject, measured from the contact of the hammer with the skin to the start of the action potential, is 47 milliseconds. Other evidence shows that of this some 35 milliseconds is nerve conduction time from the muscle to the spinal cord and back again, together with a brief central delay.

For comparison a maximal motor twitch in the extensors, elicited by an electrical stimulus to the motor nerves in the popliteal fossa, is shown above. The time base in this record is triggered by the stimulus. The latency is much shorter and the action potential and twitch tension are larger. But the duration of the action potential and the twitch are much the same as in the reflex jerk, showing that the reflex discharge itself must be a highly synchronous volley of nerve impulses. (P. A. Merton, unpublished records.)

nervous machinery controlling any muscular contraction and its function is discussed further under that heading (p. 443). The *rôle* of the stretch reflex in posture is discussed on pp. 417 and 442.

A special manifestation of the stretch reflex is the tendon jerk, of which the best known example is the knee jerk. With the subject relaxed and his knee bent the tendon below the knee cap is struck a sharp blow with a rubber-covered hammer. The extensor muscles of the knee respond with a brief twitch-like contraction, the jerk. The jerk illustrated in Fig. 14. 7 is an ankle jerk. The jerk is reflex, for it disappears if either the sensory or motor fibres to the muscles are interfered with. The latency is very short, about 20 milliseconds for the human knee jerk, little more than is needed for conduction time to the spinal cord and back. Experiments on animals show, in fact, that the central delay is less than a millisecond, time for only one synaptic junction to be passed. The afferent impulses which excite the reflex arise in the fleshy part of the muscle, not in the tendon ; striking the tendon is effective only because it causes a sudden rapid slight elongation of the muscle. The initial rapidity of stretch is the essential feature; slower stretch of larger amplitude never elicits a jerk. The initial rate of stretch needed to elicit a tendon jerk is much greater than is normally imposed on the muscles by movements of the joints, or than is used by physiologists to elicit the stretch reflex.

Several other muscles with conveniently accessible tendons exhibit tendon jerks, notably the ankle extensors (the Achilles tendon) and the biceps and triceps muscles of the upper arm. The tendon jerks are of great clinical importance because alterations in them are one of the most sensitive objective indexes available of disease of the nervous system. They disappear if any part of their reflex arc is put out of action ; and they are increased if the long motor tracts (pp. 431, 434) connecting the brain with the spinal cord are damaged, probably for the same (unknown) reasons that reflexes are augmented after complete spinal section.

More is known about the nervous mechanism of the stretch reflex than about any other reflex, but even here nothing like a complete account can yet be given and several fundamental properties still lack explanation. The receptors that excite the reflex are the muscle spindles. Muscle contains two special types of receptor ; the Golgi tendon organs lie in the tendons and in tendinous bands and aponeuroses within the muscle and respond in a straightforward way to a rise of tension whether this is due to contraction of the muscle fibres or to externally applied stretch. The muscle spindles are bundles of modified muscle fibres (intrafusal fibres) with sensory endings wrapped around a short length of the bundle (Fig. 14. 8). Impulses are set up when the sensory portion elongates. The muscle spindles lie among the ordinary (extrafusal) muscle fibres and share their attachments ; they therefore change length only when the muscle changes length. Hence they signal changes of muscle length. The muscle may shorten in length either because it contracts, in which case the tension rises, or because

the load is reduced, in which case the tension falls. Thus tension and length can alter independently in an actively contractile structure (which they cannot do in a passive structure such as a spring), so separate sense organs are needed to measure length and tension.

Impulses from spindle sensory endings travel in afferent fibres

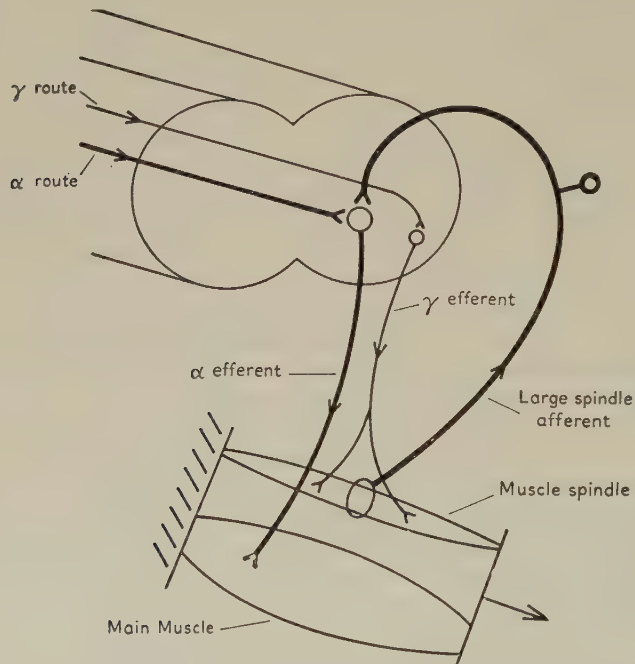


FIG. 14. 8 Diagram of the mechanism of the stretch reflex. When the muscle is pulled on, the sensory portion of the muscle spindle is stretched. As a result impulses are sent up the spindle afferent fibre, which reflexly excite the motoneurone of the muscle. Impulses are thus sent down the large motor fibre (α efferent) to the muscle and it contracts.

The sensory portion of the muscle spindle (which is itself non-contractile), is also stretched when the contractile ends of the muscle spindle contract. This they do when motor impulses reach them via their special motor nerves, the γ efferents. Contraction of the ends of the muscle spindle, due to impulses arriving along the γ route, reduces the amount of externally applied stretch that has to be applied to elicit the stretch reflex ; or it may even initiate contraction of the main muscle, via the stretch reflex pathway, without any applied stretch. (Hammond, Merton and Sutton, *British Medical Bulletin*, Vol. 12, p. 214, 1956.)

conducting with a wide range of velocities. The fast ones which are important in the stretch reflex are among the fastest in the body ; they belong to the α division of the A group (p. 372), which also includes the motor fibres. These fast spindle afferents run right through the dorsal horn of grey matter to make synaptic contact directly with the dendrites and cell bodies of the motoneurones of the muscle from which

they come, thus forming what is called the monosynaptic reflex arc. In a tendon jerk the sudden extension produced by the tap sets up a synchronous volley in the monosynaptic afferents which, when it arrives, excites a substantial fraction of the motoneurones belonging to the muscle stretched, and a twitch results.

A similar reflex occurs if the monosynaptic afferents are excited by an electrical stimulus instead of by a tendon tap. Such a *monosynaptic reflex* can seldom be obtained merely by applying electric shocks to a motor nerve, because a shock large enough to excite the monosynaptic afferents also excites the motor fibres which are in the lowest threshold, fastest conducting group too. As usual, impulses are propagated in both directions from the point of stimulation, so that as well as a volley to the muscle another passes backwards up the motor fibres to the spinal cord (an antidromic volley) at the same time as the volley goes up the monosynaptic afferents. The antidromic motor volley has the effect of inhibiting the motoneurones (antidromic block), so no reflex gets through. To avoid this in animal experiments it is conventional to cut the ventral roots so that the antidromic volley does not reach the cord, and to record the action potential of the reflex volley leaving the cord by placing electrodes on the central stump of the ventral root.

The stretch reflex also employs the fast conducting spindle afferents exciting the motoneurones monosynaptically. Thus both the tendon jerk and the stretch reflex are reflex responses to elongation of a muscle, and both employ the same reflex pathway, but with the stretch reflex an important elaboration makes itself felt, namely changes in muscle spindle activity produced by contraction of the spindles themselves. The intrafusal muscle fibres, except in the zone where the sense-endings lie, are essentially fine striated muscle fibres. They receive a motor innervation from motoneurones lying near the ordinary extrafusal motoneurones in the ventral horn, and when motor impulses reach them they contract. The sensory portion is non-contractile or less contractile, so that when the remainder of the spindle (the two "poles") contracts the sensory part is extended. This causes an increased discharge of sensory impulses, just as if the extension of the sensory ending had been due to extension of the whole muscle. (Contraction of the spindles does not cause the muscle as a whole to develop any appreciable tension, for the total cross section of intrafusal fibre is minute.) The motor nerve fibres that run to the intrafusal muscle are of smaller diameter than those going to extrafusal fibres. They belong to the γ division of the A conduction velocity group, and for this reason they are generally called "γ motor fibres"; those to the main contractile (extrafusal) muscle fibres are "α motor fibres."

Extensions of the sensory endings caused by contraction of the spindles and by stretch of the muscle are additive. It follows that, if the spindles are in a state of steady contraction owing to a stream of impulses reaching them by the γ fibres, less stretch ought to be necessary to elicit the stretch reflex than when the spindles are relaxed.

Experiment has shown that this is so. It has been found by direct recording from single α and γ motor fibres, and in other ways, that even when a muscle is quite relaxed (no α discharge) there is a continuing discharge of impulses in γ fibres. The discharge is much more intense in a rigid decerebrate cat than in a spinal cat ; the former has readily obtainable stretch reflexes, the latter little or none. The intermediate stages can be demonstrated elegantly by making use of the property of the drug procaine of blocking conduction in small nerve fibres before it blocks large ones (see also p. 375). When a drop of procaine solution of the right strength is put on a bared motor nerve it is possible slowly to block the γ fibres without affecting the α fibres. While the γ fibres to

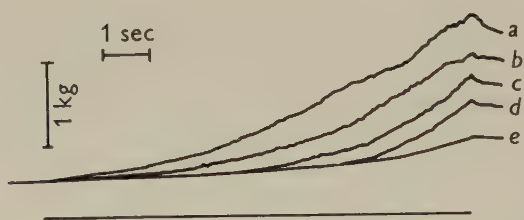


FIG. 14. 9. The effect of paralysing the muscle spindles on the stretch reflex of a decerebrate cat. The figure shows superimposed tracings of the tension developed in the soleus muscle during a series of identical extensions, each of 15 mm. at a rate of 1·5 mm./sec. The duration of each extension indicated by the signal line beneath the records. The experimental arrangements were as in Fig. 14. 6. The spindles were gradually paralysed by blocking the γ efferents with procaine. Procaine solution (0·2 per cent) was applied to the motor nerve to soleus on a piece of filter paper (see Fig. 14. 6.)

Record *a* was obtained before applying procaine ; records *b*, *c* and *d* were obtained when the nerve had been exposed to procaine for 4·5, 6·5 and 10 minutes respectively ; record *e* shows the tension developed in the passive muscle when the stretch reflex had been completely abolished by another method.

After record *d* the tetanic tension developed by the muscle to electrical stimulation of the nerve central to the procaine block was no less than it was before applying procaine. Hence the diminution in the stretch reflex was not due to blockage of α motor fibres. (P. B. C. Matthews, *Journal of Physiology*, vol. 147, p. 547, 1959.)

an extensor muscle of a decerebrate cat are being blocked in this way the extension required to elicit a stretch reflex increases progressively as block deepens (Fig. 14. 9). The spindles are relaxing and a greater stretch is needed to produce the same rate of spindle discharge.

The degree of contraction of the spindles is said to control the " bias " of the stretch reflex. It is suspected that much of the unresponsiveness in spinal shock may be due to lack of bias, " spindle paralysis," due to a cutting off of excitation from the higher parts of the nervous system to the γ motoneurones ; conversely, decerebrate rigidity is thought to be associated with " spindle cramp ". Whether the increased stretch reflexes and tendon jerks found in muscles rendered spastic by

disease of the long motor tracts in man are also due to increased spindle bias is not certain ; there is some evidence that they are.

Although the stretch reflex and the tendon jerks are both exaggerated in decerebrate rigidity in animals or in spastic human muscles, and are both diminished if any part of the reflex arc (*e.g.* the dorsal roots) is damaged, they do not always behave similarly and that is why it is best to treat them separately, even though they share the same reflex pathway. In spinal shock, in light barbiturate anæsthesia, and after damage to the cerebellum (p. 439) tendon jerks of normal briskness may be present when the stretch reflex cannot be obtained at all. Why this should be is not understood ; in all these conditions the loss of stretch reflex is probably due, at least in part, to diminished spindle bias, but relaxation of the spindles may have less effect on their response to a sudden jar. In addition there is evidence that the spinal cord may continue to respond to artificially synchronised volleys at a time when normal maintained responses to impulses arriving in an asynchronous stream are lost.

Another uncertainty concerns the part played by the Golgi tendon organs in the stretch reflex. Various pieces of evidence show that impulses from the tendon organs inhibit the motoneurones of the muscle from which they come. If the knee of a spastic man is forcibly bent at first the exaggerated stretch reflex resists powerfully, but at a certain point resistance collapses. This is called the " clasp-knife " response. A similar reflex response in chronic spinal or decerebrate animals is called the lengthening reaction ; it is attributed to impulses from the Golgi tendon organs. There is some evidence that the inhibition acts on the γ motoneurones causing a sudden relaxation of the muscle spindles. The significance of the lengthening reaction is not clear ; it is sometimes said to be protective against excessive tension, but this (as Sherrington himself saw) it clearly is not, for it may occur at tensions which must often be exceeded in ordinary life and are obviously not dangerous. The lengthening reaction has an inverse, the shortening reaction ; if the bent knee is straightened out it is again found to resist bending ; in other words the muscles have gone back to the initial state and the stretch reflex and lengthening reaction can be elicited once more. Much is obscure about the mechanism and significance of the lengthening and shortening reactions, but it is clear they have the tendency to make a limb stay put in whatever position it is made to adopt, in Sherrington's word, they render posture " plastic."

Posture and Locomotion

As already emphasised it is unrewarding to think of posture and locomotion as reflex responses ; how they are carried out is the interesting question, largely unanswered it is hardly necessary to add. The basic mechanisms for both standing and walking lie in the spinal cord, for spinal dogs and cats if allowed enough time to recover from shock will stand, and the hind limbs supported clear of the ground will perform indubitable walking movements in response to such stimuli as a pinch

on the buttocks. Spinal standing and walking are imperfect performances; in order to stand and walk in a normal manner a cat or dog requires not only the spinal cord but the brain stem up to the front of the midbrain. In an ordinary rigid decerebrate preparation the midbrain is sectioned between the superior and inferior colliculi. If the section is in front of the superior colliculus, so as to leave behind the red nuclei (Fig. 14. 16), decerebrate rigidity does not result and the preparation (known as a midbrain animal) stands with a normal distribution of contraction between the extensors and flexors of the limbs. Why this should be is not known; decerebrate rigidity is discussed further on p. 444.

The brain stem is also necessary because it is the site of various reflexes that keep the animal balanced on its legs and restore it if it is thrown off balance. A spinal animal falls over very readily. Little is known about the mechanism of these righting reflexes except the sense organs they employ. In the midbrain animal the most important are the vestibular balance organs in the labyrinth of the inner ear (p. 513); they reflexly cause the head to be kept upright. The body reflexly follows the head, but for this to occur the angle between head and body must be known; this is signalled by joint receptors in the neck vertebra. With both labyrinths destroyed a midbrain animal can still right itself using information from pressure receptors in its flanks. In most intact animals the eyes are of at least as much importance as the balance organs for righting; the rabbit is an exception: its eyes are of no help.

Man has a more difficult balancing task than four-footed animals, particularly if he stands with his feet together; he uses his eyes and differential pressure sensations from the soles of his feet. If sensation in the feet is impaired (as it is for instance in alcoholic neuritis) the patient falls over if he stands with his feet together and then shuts his eyes (Romberg's sign). The labyrinths do not provide sufficiently sensitive information to enable him to balance; in fact in man the main function of the labyrinths is probably to steady the eyes, to produce compensatory movements of the eyes so that fixation of external objects can be automatically held when the head turns. A man is little disabled by the loss of labyrinthine sense. He can even ride a bicycle.

Apart from the general brain stem reflexes concerned with balance there are also individual mechanisms in each limb for throwing the muscles into an appropriate state of contraction when an animal is placed on its feet. Thus when a midbrain animal is held up in the air its limbs hang limply, but as soon as a paw makes contact with the ground the whole limb stiffens into a standing posture; both flexor and extensor muscles contract to fix the joints. This is called the positive supporting reaction. The stimuli immediately responsible are pressure on the pads of the foot and stretch of the muscles to the digits and feet. The positive supporting reaction, however, is by no means just a stretch reflex, for many of the muscles that contract are not stretched at all.

The importance of the stretch reflex in standing has probably been overemphasised in the past. The attitude of a decerebrate cat was interpreted as an exaggeration of normal standing, and the fact that the extensor rigidity is due to greatly augmented stretch reflexes suggested that normal standing could be accounted for in terms of stretch reflexes of a more appropriate vigour. This argument has lost its force now it is known that there is another form of decerebrate rigidity, in which the attitude of the animal is the same, but which does not depend on the stretch reflex (p. 444). Furthermore decerebrate rigidity differs significantly from normal standing in that rigidity involves only extensors whereas both extensors and flexors contract in standing. The truth may prove to be that internal motor mechanisms, so far unknown, cause the appropriate muscles to contract and, as in many other actions (see p. 443), the contracted muscles exhibit stretch reflexes. The stretch reflex is obviously well adapted to oppose the forces that gravity applies to the muscles used in standing. On this view stretch reflexes remain of importance in standing but standing involves much more than the existence of stretch reflexes in extensor muscles.

For locomotion there is better evidence that the mechanism that produces the contractions of the flexors and extensors of a limb (in this case in sequence), does not depend on reflexes from the limb. A chronic spinal cat with all dorsal roots severed except those belonging to the tail, will still make walking movements of the legs when the tail is pinched. This strongly suggests that the organisation of walking is internal, and that walking is not of the nature of a chain reflex, in which movement of a limb forward sets up a reflex to carry it back and so on. Reflexes, however, are no doubt of great importance in modifying and making precise the movements of walking, although definite evidence as to how exactly they do it is lacking.

Voluntary Action

A voluntary action like picking up a pencil involves first seeing and recognising the pencil and then making the appropriate movement to take hold of it. This description, it will be seen, is of the same form as the description of reflex action that was arrived at in a previous section. Quite possibly it glosses over some essential difference between voluntary and reflex action, but it takes a good deal of the mystery out of both to emphasise their formal similarity. Of course, voluntary actions are often immensely more complex than reflex actions, and they are attended by consciousness. Physiology has nothing to say about the nature of consciousness except that it is something that seems to accompany the most elaborate reactions of an animal to its environment. It may mean that some fundamentally different kind of process is occurring or it may not. Whichever is the case it certainly makes voluntary actions, in spite of their complexity, easier to think about than reflexes because we have an immediate subjective awareness of the steps involved in a voluntary action (recognition, the decision to

make a particular movement, etc.) that can only be reached by indirect argument in the case of a reflex.

In voluntary action the recognition of what is going on in our surroundings and the decision what to do about it both take into account an enormous store of memories of what the messages from our sense organs signified in the past and what the results of previous muscular efforts were. When a large number of factors have to be weighed up simultaneously in this way the brain seems to choose to do the job in a thin sheet of nerve cells the area of which, rather than the thickness, increases as the complexity of the problem increases. Such sheets of grey matter develop on the surface of the brain and are hence called "cortex," from the Latin word meaning outer shell. Conscious voluntary activity takes place, mainly at any rate, in the cerebral cortex ; as the old writers put it, it is the organ of mind. In ascending the evolutionary scale the area of the cerebral cortex increases in step with increasing complexity of behaviour, ending with the enormous cerebral cortex of man.

The cerebral cortex has tackled its manifold tasks by subdividing and spreading them out over its surface, the more complex the task the larger the area devoted to it. In man there are large areas devoted to vision, hearing, muscular movement, sensations from the skin surface, etc. In picking up a pencil, there is evidence that seeing the pencil and recognising it are associated with activity in and around the visual area of the cortex, and that the movements of the arm to pick it up are the responsibility of the area concerned with movement (the motor area). Thus it would appear that the steps into which we subjectively divide the action of picking up the pencil are in fact related to nervous processes occurring in distinct regions of the cerebral cortex. That this is so has only become clear within the last hundred years but already the facts of functional localisation in the cortex are so familiar that it is easy to forget that they are not merely an anatomical matter. Just as the idea of separate motor and sensory nerves is now taken so much for granted that we are apt to forget that their discovery was not at the time just a matter of the anatomy of the spinal roots, so too we tend to think that modern neurology has merely answered the anatomical questions, where are functions A, B and C localised in the cortex? But a hundred years ago it was not known what the functions of the cortex were, nor whether they were generalised or localised, so the questions could not have been asked.

Thus the prime importance of the observations which showed a localisation of cortical function was to reveal what kinds of function the cortex performs. The story, indeed, is by no means complete yet ; there are large areas of which we only know in the most general way what they do, and some functions (*e.g.* vision) seem to be much more definitely localised than others (*e.g.* memory, which some people think is not really localised at all, see p. 446), but the main outlines seem clear enough. Roughly speaking the back part of the cerebral cortex is sensory in function and the front part motor, the dividing line being

the central sulcus. In the back half of the cortex are areas devoted to vision, hearing, smell and taste, and to sensations from the skin and the body generally (somatic sensations). These areas (except smell and taste) are shown in Fig. 14. 3. They are called the primary or receiving areas because impulses from the respective sense organs are actually brought by nerve fibres to these areas. The cortex around and between the receiving areas seems to be concerned with interpreting what comes in, with building up a single picture of the external world from the information supplied by all the senses and with assessing the significance of what is going on. One highly specialised aspect of this general function with which special areas are associated is the comprehension of speech and the written word.

The front half of the cortex is concerned with the execution of motor acts and in a general way with planning. The detailed orders to the muscles to perform particular movements are sent out by the cortex just in front of the central sulcus. Again speech gets special treatment ; its motor organisation occupies a patch of cortex (Broca's area) in front of the main motor area.

The motor area and the four sensory areas (somatic, visual, auditory and olfactory) are the only parts of the cortex which have large and obvious tracts of nerve fibres connecting them to other parts of the nervous system ; for this reason they are often referred to collectively as projection areas ; the parts in between, which are principally connected to other parts of the cortex, are called association areas ; formerly they were called silent areas because electrical stimulation or surgical removal of these areas, particularly in animals, is often without obvious effect. In lower mammals the projection areas occupy a large part of the cortex but in man the proportion is quite small. A man's eyes, ears and hands are not much better than those of a chimpanzee, nor are the absolute sizes of the cortical areas to which they project greatly different ; he is superior chiefly because of the far greater use he makes of the information from his sense organs and the far greater skill with which he uses his hands ; this he owes to his enormously expanded association areas.

In man, and indeed in nearly all vertebrates, the most important source of information about the environment is vision. What we see occupies so much of our attention that with the great majority of mankind it can be assumed that if their eyes are shut they will not be attending to anything ; conversely, though we can, by concentrating, disregard many sounds, smells and bodily sensations, it is very difficult indeed not to pay attention to what is passing before our eyes. The dominance of vision makes it natural to begin the more detailed discussion of the cerebral cortex with vision, more particularly since the anatomical peculiarities of the visual pathway have effects which are felt throughout the whole of the nervous system.

The Visual System. In man it is established with great certainty that damage to the occipital lobes causes blindness, and that lesions (*i.e.* areas of damage) elsewhere in the cerebral cortex never have this

result. Hence we speak of a visual sensory area in the occipital lobes. When a lesion is small and circumscribed, blindness is limited to a particular part of the visual field. This result could not have been foreseen ; for supposing, speaking metaphorically, the visual cortex had functioned like a lens to focus nerve impulses from the retina on to consciousness, then putting out of action one part of it would have caused a general dimming of the whole field and it would not have mattered which part was involved, only the total area would count. There may be functions of the brain, such as memory, in which the same disability might be produced by lesions in different areas of the

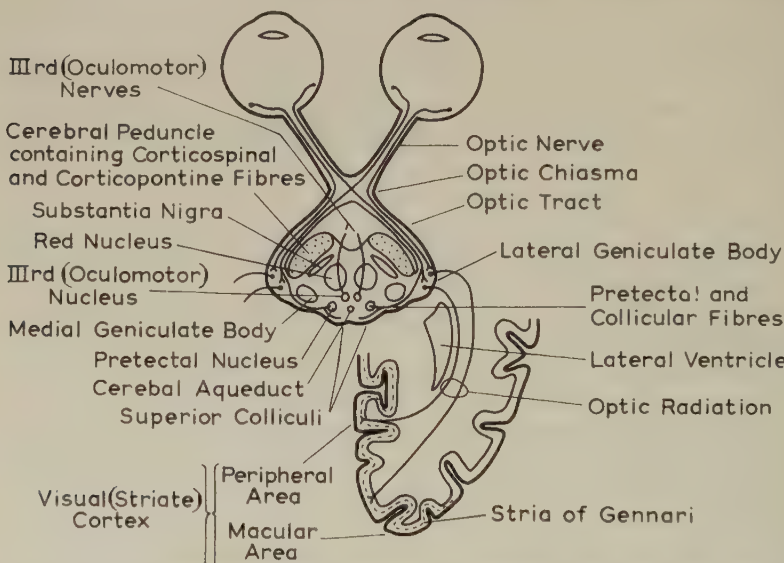


Fig. 14. 10. The human visual pathways. Horizontal sections, roughly to scale, are shown of the right occipital lobe and of the upper part of the midbrain adjoining the thalamus. Some non-visual structures in the midbrain are also labelled.

The pretectal and collicular fibres are separate optic nerve fibres and not branches of geniculate fibres as the diagram has it.

cortex in this way and in these cases it is sometimes written that they are functions of the cortex " as a whole." We shall return to the question later (p. 446) but vision at any rate is very clearly not like that ; blindness only results from cortical lesions if they are in the occipital lobe and blindness in a certain part of the field of vision can only be produced by a lesion in a certain area within the visual cortex.

Optic nerve fibres from the retina do not run direct to the visual cortex but go to the lateral geniculate bodies, two nuclei lying on the under side of each thalamus at the base of the cerebral hemispheres ; from the neurones in them large tracts of nerve fibres arise, called the optic radiations, which curve backwards to the occipital cortex (Fig. 14. 10). Now over most of the cerebral hemispheres the cortical grey

matter is remarkably uniform in naked eye appearance, consisting of a featureless sheet some 2 mm. thick, but the area to which the visual radiation runs is distinguished by a conspicuous white band (of nerve fibres) in the middle of the grey matter, called the stria of Gennari ; hence the visual area is also known as the striate cortex. The function of the stria is unknown. In man, part of the striate area is on the tip of the occipital pole but the greater part is out of sight in the deep groove where the medial surfaces of the two hemispheres face each other.

By no means all optic nerve fibres go to the lateral geniculate bodies and thence to the cortex. In lower animals like fish, frogs and birds the main visual pathway runs to the optic lobes on the dorsal side of the midbrain, and even in man 20–30 per cent of all fibres from the retina go past the geniculate bodies and end in the same region of the midbrain, known in higher animals as the superior colliculi (superior corpora quadrigemina), or in the grey matter deep to the colliculi (pretectal nuclei). Animal experiments indicate that the pretectal fibres are responsible for the contraction of the pupil when a light is shone in the eye, but there is no evidence about the function of the very numerous collicular fibres in man. It is a fair surmise that they are involved in automatic control of eye movements (Chapter 18, p. 562) but only a surmise.

Even in primitive vertebrates visual impulses from external objects to the right hand side of an animal are fed to the left half of the brain and vice versa. The reason for this crossing over, or *decussation*, is not properly understood although it is often thought to be a consequence of the reversal of the optical image caused by the camera-like structure of the vertebrate eye with its single lens. To remember the scheme on which the different parts of the retinæ are projected on to the human visual cortex it is convenient to think first of an imaginary man with a single central eye. The natural method seems to be for the upper part of the retina to be connected to the upper part of the visual area and the lower part to the lower part ; the right hand side to the right hand half of the visual area, *i.e.* to the right occipital pole ; and the left hand side to the left occipital pole. The two real eyes both see almost the same field and each has the same connections as the hypothetical single eye, which means that fibres from the nasal halves of each retina have to cross in the optic chiasma. As a matter of fact things are not quite so simple because although it is true that fibres from the right hand halves of both retinæ go to the right hand occipital cortex, with upper retinal fibres keeping to the top and lower to the bottom, a reversal has occurred such that the central part of the field (the macula) is not represented medially as it ought to be, if the above scheme were carried through, but laterally ; and the peripheral retina is represented medially, *i.e.* on the opposed surface of the hemispheres. Thus if we imagine an object moving from left to right in front of a man who gazes steadily straight ahead, the area of cortical activity representing the object starts on the right striate cortex deep in the groove where it

faces its fellow ; it then moves posteriorly and laterally on to the tip of the occipital pole ; when the object is dead ahead the active area jumps over to the opposite pole and moves medially and anteriorly down on to the medial surface again.

Part of the evidence for the above statements is that division of one optic tract, or removal of the whole of one occipital lobe in man, results in complete blindness of the opposite half of the visual field of both eyes (a hemianopia). The vertical division of the visual field between left and right cerebral hemispheres revealed in this way seems to be quite sharp and splits even the macula, the part of the retina we use when looking straight at an object, straight down the middle. It is perhaps rather surprising that, when a small object moves from one side to the other of the central line and its area of cortical representation jumps from one hemisphere to the other, there is no trace of discontinuity in our sensation. Of course corresponding areas of the two hemispheres are connected together by enormous numbers of nerve fibres running in the corpus callosum (Fig. 14. 3), and these connections may be important in welding the two visual areas into a functional whole. But division of the whole corpus callosum has been performed several times in man with remarkably little effect at all, and in particular with no conspicuous visual symptoms. It is one of the minor mysteries of cerebral physiology that whereas the mind is single the cerebral hemispheres are a double organ (not all their functions, however, are doubled as we shall see (p. 435)).

We have seen that the retinæ are, as it were, mapped on to the visual cortex, but nothing has been said as to scale. Here we come on the most notable instance of the general rule that the area of cortex devoted to a part of the body depends on its importance and not on its size. The total area of visual cortex is not far different from the area of the two retinæ but something like half the total area of visual cortex is probably devoted to the macula, the central part of the retina about 1 mm. in diameter where visual acuity is highest. Thus the macula occupies a very much larger area on the cortex than it does on the retina.

When a lesion of the striate cortex causes a scotoma (a localised area of blindness) vision is lost as a whole in that area ; it never happens, for instance, that colour vision is lost selectively, or that threshold to perception of light remains unaltered but objects cannot be perceived ; all these functions are affected together, although in partially damaged areas not necessarily to the same degree. Scotomas can often only be found by careful examination, for patients are frequently unaware of blindness of cortical origin and may deny that they have any trouble in seeing when they are grossly disabled. It is a general characteristic of cortical lesions, which goes under the name of anosognosia, that patients may lack insight into their condition and may not complain of their disabilities. This may seem less odd if we reflect that no one would expect a man who had had the whole of his cerebral cortex removed to realise how stupid he was. To remember what it was like

to be intelligent he would have to be intelligent still. Similarly the visual cortex is not only the part of the cortex that does the seeing ; it seems also to be the part that knows what seeing is.

The parts of the occipital lobe around the striate area and the posterior part of the parietal lobe adjoining it appear to be concerned with interpreting what we see with the striate cortex. The symptoms of disease in this region are very various and fascinating, but there is only space to mention a few. To start with one that has been mentioned already, damage to the posterior part of the parietal lobe may abolish the reflex blink to a threatening gesture. Such a patient may have no detectable interference with vision and certainly sees a fist approaching, but his brain fails to interpret it as a threat, at any rate in time to do anything about it. Other lesions may upset judgement of distances. The patient may state that he clearly sees a chair placed in front of him, and yet when invited to step forward he walks straight into it and appears surprised to find it there. In these cases the disability may not be confined to a failure to judge distance, but may extend to a general inability to grasp the spatial relations of objects to each other and to himself. He may be unable to find his way about a house he has lived in for years. If five pennies are put on the table before him he cannot count them because, as it were, he never knows where he has got to. Another manifestation of disease in this region is the failure to recognise things which nevertheless are seen clearly enough. The patient when shown various common objects such as a box of matches or a pair of scissors, etc., may not be able to put names to them or describe what they are used for, although he can do so at once when allowed to handle them. The frequently associated inability to read will be discussed later under the subject of speech.

The Auditory and Olfactory Pathways. The auditory area of the cortex is in the temporal lobe, in man in the part opposite the bottom end of the central sulcus (Fig. 14. 3). Impulses from the ears reach the auditory cortex via the VIIIth nerve nucleus in the medulla and the medial geniculate body. In addition to this geniculate pathway there are a large number of fibres that run to nuclei in the medulla and mid-brain, notably to the inferior colliculus. The automatic turning of the head and eyes towards a source of sound is suspected to be one function of these pathways (compare the visual fibres to the superior colliculus, p. 421). The geniculate pathway is mainly crossed but a substantial fraction is not ; each geniculate body and hence each auditory cortex is in connection with both ears. The result is that, in contrast to vision, damage to one auditory area or to the auditory pathway on one side does not cause marked deafness. As lesions large enough to affect both pathways are likely to be fatal, deafness is a rare symptom of disease of the brain. Because of this much less is known about hearing than about vision, even the position of the auditory area is not clearly delimited in man.

Electrical recording from the exposed cerebral cortex of the anæsthetised dog has shown that impulses excited by notes of different fre-

quencies go to different points on the auditory cortex. The frequencies are spread out along a strip, high notes in front and low notes behind. Each octave occupies about 2 mm. of the strip. In man a jump of an octave in pitch gives the same subjective impression in whatever part of the scale it occurs. The frequency mapping in man may be quite different from that in the dog, but it is certainly very suggestive that in the dog a jump of an octave always moves the point of arrival of impulses the same distance along the cortex.

Signals from the balance organs of the inner ear play little part in conscious life, unless powerful stimulation makes us giddy or sea-sick (see also p. 416). The cortical area dealing with vestibular sense appears to be in the temporal lobe near the auditory area.

The parts of the cortex concerned with smell (and taste) are hidden away underneath the cerebral hemispheres. The olfactory nerves go to the olfactory bulbs under the frontal lobes and these are connected to the olfactory area of the cortex in the uncinate region under the temporal lobe. Smell is a relatively unimportant sense in man, and the olfactory parts of the brain are poorly developed. It still has survival value by occasionally giving an invaluable danger signal of a fire or a gas-leak, but loss of smell is a small disability compared with the loss of any of the other senses, although enjoyment of food is much diminished for it depends largely on the sense of smell. Uncinate epilepsy is mentioned later (p. 446).

Somatic Sensation. Afferent nerve fibres from sense-endings in the skin and the interior of the body are gathered together into peripheral nerve trunks and run to the spinal cord via the dorsal roots. The cell bodies of these nerve fibres lie in ganglia on the dorsal roots (Fig. 14. 1). These particular cell bodies are merely trophic centres ; nothing of nervous interest happens in a dorsal root ganglion, impulses just pass through as if it were an ordinary piece of nerve. On entering the cord afferent fibres divide into ascending and descending branches ; from these branches arise collateral branches that run into the grey matter for reflex purposes.

Ascending branches that carry messages destined for the brain, to arouse conscious sensation, do one of two things (Figs. 14. 11 and 14. 12), either they immediately enter the dorsal column (on the same side they enter) and run straight up to the dorsal column nuclei in the medulla, or, after ascending for a segment or two, they terminate by making synaptic contact with nerve cells in the grey matter of the dorsal horn. The axons of these cells run all the way to the thalamus, the main sensory nucleus of the brain, which lies near the midline at the base of the cerebral hemispheres ; to get there they cross the midline and run up the opposite side of the spinal cord in the ventro-lateral column of white matter. The tract is called the spinothalamic tract, but it is not a clearly demarcated area of white matter like the tracts in the dorsal columns. The fibres composing it are mixed in among a much larger number of descending fibres and other ascending fibres.

The axons that arise from the cells of the dorsal column nuclei in the

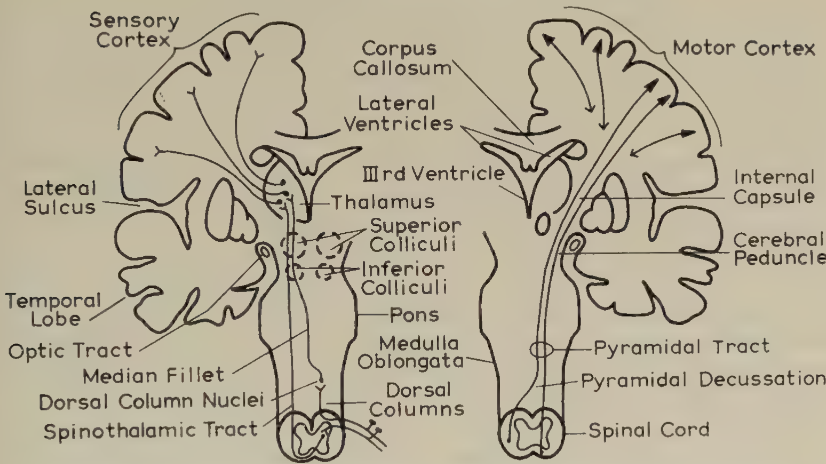


FIG. 14. 11. Diagrammatic Coronal (frontal) sections of the human brain to show the sensory and motor pathways. Plane of section indicated on Fig. 14. 3. Roughly to scale except that the diameter of the brain stem and spinal cord is exaggerated. Some structures, shown here but not labelled, are labelled on Fig. 14. 16. For spinal tracts see also Fig. 14. 12. The outlines of the colliculi on the dorsal surface of the mid-brain shown dotted. The median fillet is also known as the medial lemniscus.

medulla also cross the midline and run to the opposite thalamus. In the medulla they form a conspicuous slab-like band of white matter called the median fillet. Thus all nerve impulses that give rise to conscious sensation find their way to the opposite thalamus ; either they run up the dorsal column on the same side as they enter the cord and cross over immediately above the dorsal column nuclei, or they cross within a few segments of entering the cord and run up the opposite spinothalamic tract. These facts are important because the two pathways do not carry the same types of sensation. The spinothalamic pathway conveys sensations of pain, temperature and crudely localised mechanical contact. The dorsal columns are responsible for all the more refined types of touch sensation, such as we possess pre-eminently in

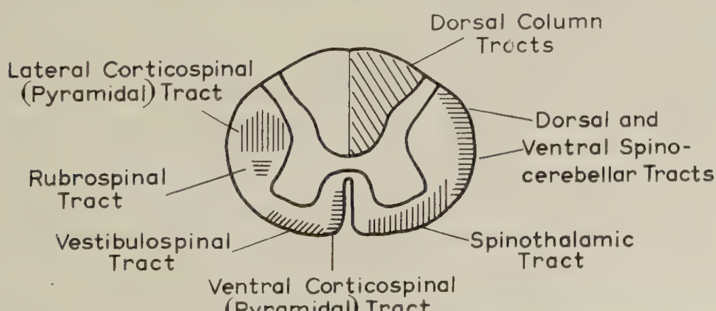


FIG. 14. 12. The approximate positions of the principal tracts in the human spinal cord, ascending tracts on the right, descending tracts on the left.

our finger tips, and for sensations of position and movement of the limbs.

The result of this segregation of sensory pathways is that lesions of the spinal cord other than complete division seldom cause a complete loss of all forms of sensation in any part of the body, such as occurs when a peripheral nerve is divided. As a rule spinal disease causes a *dissociated sensory loss* in which some forms of sensation are lost and some retained. The clearest instance is when injury or disease sections either the right or the left half of the spinal cord, giving rise to what is known as Brown-Séquard's syndrome. Below the level of hemisection the patient suffers a loss of dorsal column sensation on the side of the lesion and a loss of spinothalamic sensation on the opposite side.

The dorsal column loss shows itself by an inability to localise contacts ; if one of his fingers is touched the patient will feel it but be unable to say which finger it is ; again the ability to distinguish two points from one when the tips of a pair of compasses are pressed on the skin is greatly impaired, and the points may have to be opened out several inches before they are recognised as two. Position sense is lost and the patient cannot say what position his joints are in or whether they are being flexed or extended by an examiner. The loss of discriminative tactile sense and the loss of position sense together prevent him from recognising the shape and texture of objects which are put in his affected hand ; this important symptom is called astereognosis and it makes the affected hand useless for most purposes.

The spinothalamic loss on the opposite side of the body makes the patient insensitive to pain, however aroused, and incapable of distinguishing hot from cold objects. No straightforward tactile loss is detectable for dorsal column sensation remains on that side but the loss of the spinothalamic type of tactile sensibility can sometimes be detected by the absence of sensations of itching and tickling. Itch and tickle are curious quasi-painful sensations that are believed to be carried in the spinothalamic tracts. The principal symptoms, therefore, are astereognosis and loss of position sense on the side of the section and loss of pain and temperature sensation on the opposite side—a dissociated loss. Simple contact, *e.g.* a brush with a piece of cotton wool, is felt everywhere because it can travel by either the direct or the crossed pathway.

Both the ascending sensory pathways eventually end in the ventral part of the thalamus ; from these a large tract of fibres (forming the posterior part of the internal capsule) runs to the strip of cerebral cortex behind the central sulcus, the (somatic) sensory area (Fig. 14. 3). Destruction of the sensory area in man causes a loss of sensibility over the opposite half of the body ; but, whereas loss of one occipital lobe gives rise to permanent complete blindness in the crossed half-field (p. 422), even very extensive damage in the sensory area does not cause permanent complete anæsthesia on the opposite side of the body. Except, perhaps, for a day or two after an acute lesion such patients

are found still to appreciate pain, differences of temperature, and ill-localised contact on the opposite side of the body. The interpretation of this phenomenon is controversial but the simplest view, originally put forward by Head and Holmes, is that consciousness of these sensations depends in some way on activity in the thalamus. All the more highly developed forms of sensation require the integrity of the sensory area of the cortex. Roughly speaking, sensations associated with the thalamus are of the type carried in the spinothalamic tracts, and sensations permanently lost after cortical damage are those carried in the dorsal columns. The former tend to be associated with feelings of pleasure or of unpleasantness, warmth, cold and pain (see pp. 433, 436 and 460 for further evidence that the central parts of the brain are concerned in emotion) ; but cortical sensations, such as distinguishing two compass points from one, are without emotional colour.

Lesions of the thalamus itself, besides causing a loss of sensation, may lead (after an interval of some weeks) to spontaneous and intractable pain, referred to the opposite side of the body ; or it may happen that, without spontaneous pain, a stimulus such as a light pin-prick may cause disproportionately severe pain. Such symptoms, known together as the thalamic syndrome, never, or almost never, occur with lesions of the sensory cortex (or indeed with lesions anywhere else in the brain), which is another reason for thinking that the thalamus is particularly concerned in appreciation of pain.

Lesions localised to a part of the sensory area cause loss of sensation in a particular part of the body in the same way that small lesions of the occipital cortex cause a localised patch of blindness in the visual field. In man the extent of the sensory area and the way the body is represented on it have been learnt mainly from the results of local damage, and by stimulation of the exposed cortex at operation. The brain substance itself and its immediate coverings are insentient, so that operations on the brain can be performed with only a local anæsthetic to insensitise the skin. It is then a simple matter to stimulate a point on the brain electrically by passing a low voltage 50 cycles alternating current between a round-tipped wire electrode resting on the cortex and an indifferent plate electrode on the skin elsewhere. When the electrode is on the post-central cortex the patient has a sensation which he feels in some part of the opposite side of his body. The electrical stimulus presumably does not imitate the normal pattern of activity caused by sensory impulses arriving at the cortex, for the sensation is indefinite and unfamiliar, not unlike the sensation from electrical stimulation of a cutaneous nerve (p. 400), but there is no reason to doubt that it indicates which part of the body normally sends sensory impulses to the point being excited. (With strong stimulation the referred sensations may become painful, an observation which suggests that the cortex is not wholly unconcerned in the appreciation of pain.) In this way it has been found that the body is mapped upside down on the sensory area ; messages from the foot and leg go to the top of post-central cortex, the middle region is concerned with the hand and

arm, and the bottom with the face. As with vision the scale of mapping is very uneven ; important parts, like the fingers, have a large area devoted to them, the skin of the back a very small area.

In animals the sensory area can be mapped by recording electrically the places which receive impulses when particular parts of the body are touched. The animal has to be fairly deeply anæsthetised to cut down the noisy background of spontaneous activity in the cortex. Then a fine wire inserted into the cortex will detect sensory impulses arriving. Interestingly enough although they have crossed at least two synaptic junctions (in the dorsal column nuclei and the thalamus) the trains of impulses from the sense-endings are little modified and what arrives

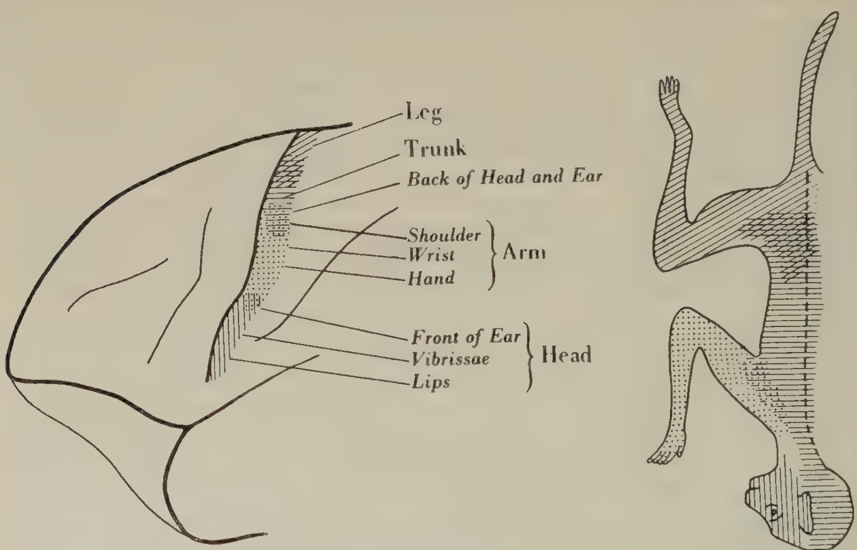


FIG. 14. 13. The somatic sensory area in a rhesus monkey, showing the large area devoted to the hand and face. The trunk region includes the back of the neck and head. (Adrian, *Journal of Physiology*, vol. 100, p. 159, 1941.)

at the cortex is very much like what can be recorded from sensory fibres in peripheral nerve. The difference is that with parts of the body, such as the back, in which the cortex is not much interested, several or many sense-endings may share a single projection fibre, so that touches at widely separated points, supplied by different spinal nerves, all cause activity in the same cortical fibre. Fig. 14. 13 shows Adrian's map of the sensory area in the monkey, which is not unlike that of a man.

In other animals the rule that only important matters get a hearing in the cortex is sometimes even more obvious than in the monkey and man. A pig uses its snout not only for finding its food but also for digging it up ; it is a pig's principal executive organ ; the legs are little more than props for the body. Adrian found that the snout, so far as

he could tell, was the only part of the body that sent sensory impulses to the cortex. The whole somatic sensory area was snout (Fig. 14. 14). And the absolute area of cortex devoted to the pig's snout was larger relative to its size than the area of human cortex concerned with the hand. Other mammals also devote much of the sensory area to the nose and lips, particularly when they bear vibrissæ (the long hairs with which cats and rats feel their way in the dark), but if the legs are used for fighting or digging they also are represented.

As regards the crossing over of sensory pathways by which sensory impulses from one side of the body are carried to the thalamus and cortex of the opposite side, the accepted explanation is that this occurs in order that they should be dealt with on the same side of the brain that deals with vision for their side of the body. The mammalian brain is dominated by vision, and because the visual representation of

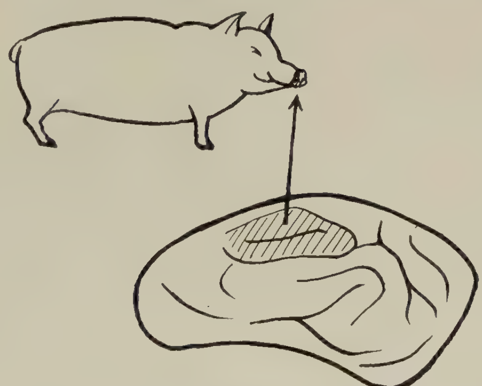


FIG. 14. 14. The somatic sensory area in the pig. The whole area is devoted to the opposite half of the snout. (Adrian, *Brain*, vol. 66, p. 89, 1943.)

external objects on the cortex is crossed (the left striate area dealing with the right half of the visual field and so on) everything that has to be correlated with vision crosses too ; that includes not only somatic sensation but hearing and the motor pathways too. It lends support to the correctness of this view that when, exceptionally, somatic sensations are to be correlated with olfaction, whose pathway does not decussate, the somatic pathway remains uncrossed too. Thus Adrian found in ruminants (the sheep and goat) in whom smell is probably the main guide to what is to be eaten next, that the sensory area for the lips was uncrossed. The remainder of the body had a crossed representation as usual. In the cat and in other animals in which vision is used in feeding the cortical representation of the lips was crossed.

In man the symptoms of damage to the somatic sensory area, say the region concerned with the hand, are in the first place loss of all delicate, discriminative skin sensations by which objects are recognised

and fine manipulations directed. The patient also tends to be inattentive to the affected hand, so that if quantitative testing of sensation is attempted, performance is irregular. If touched simultaneously on both hands he may fail to notice the touch on the affected hand, although he does so all right if it is touched alone. It seems that the affected hand tends to slip out of his mind ; indeed with large lesions it may go out of his mind altogether, giving rise to the bizarre symptom that the patient when shown the affected hand denies that it belongs to him. He has apparently lost the part of the brain that deals with his conception of his hand. This is analogous to the patient with damage to the occipital lobes who is not aware that he cannot see. Needless to say a severe cortical sensory loss renders the hand useless, and, although it is not paralysed, to an onlooker it may appear to be, for the patient attempts nothing with it.

The Voluntary Motor System. The orders that the cerebral cortex issues to the muscles of the body leave from a strip of cortex in front of the central sulcus called the motor area (Fig. 14. 3). If the motor area is damaged or removed the patient develops a voluntary paralysis on the opposite side of the body. He knows what he wants to do but when he tries to move the affected parts nothing happens. Lesions elsewhere in the cortex never have this result ; this is the first reason for regarding the precentral cortex as an area specially devoted to voluntary movement.

Movements of different parts of the body are dealt with by different parts of the motor area ; the arrangement matches that in the sensory area on the opposite side of the central sulcus, leg and foot at the top, arm and hand in the middle and face at the bottom. Fig. 14. 15 shows the motor area of a chimpanzee. The human motor map is known to be very much the same but a detailed map of a whole healthy human motor area is not available. As in the sensory cortex there are large areas for important parts like the hand and only small areas for parts like the buttocks which have little to do directly with voluntary activities. Since on the whole important parts are small and large parts unimportant it turns out that roughly speaking the area of cortex devoted to a part of the body is inversely proportional to its bulk. It is presumably because sensation, particularly touch, is of such great importance in guiding movement that the cortical area which controls the movement of a part lies next door to the area to which it sends its sensory messages. In fact the motor and sensory areas to some extent run into each other, for when stimulating points on the exposed human cortex electrically it is found that sensations (referred to the opposite side of the body) are sometimes felt when the electrode is on the precentral cortex, and movement sometimes results when the electrode is on the postcentral cortex.

Movements of the opposite side of the body elicited by electrical stimulation are a hall-mark of the motor area. The ease with which movements can be obtained by electrical stimulation of the motor area, compared with the impossibility of obtaining movements from most

other places in the cortex, is the second reason for linking the precentral area with the control of voluntary movement. The movements occur in the same part of the body which would become paralysed if the region of cortex around the stimulating electrode were excised. Thus the way in which the parts of the body are mapped on to the motor cortex can be discovered by electrical stimulation.

The motor area is connected to the spinal cord by numerous fibres running in the corticospinal or pyramidal tract (Fig. 14. 11). The

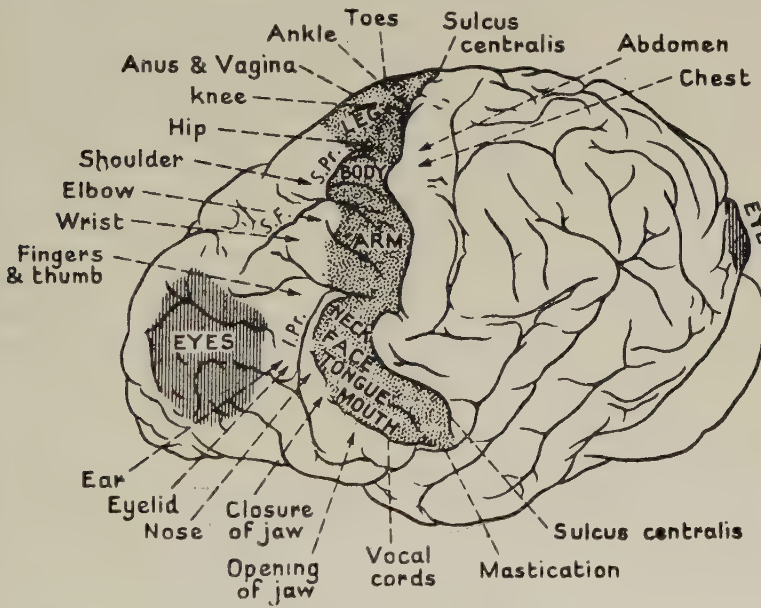


FIG. 14. 15. The motor area of a chimpanzee mapped by electrical stimulation. There is much overlapping of the regions and their subdivisions which the diagram does not attempt to indicate.

Conjugate eye movements can be obtained by stimulating the visual area in the occipital lobes and also from the frontal lobe in front of the motor area. These movements (not mentioned in the text) differ in their characteristics from movements obtained from the ordinary precentral motor area. (Sherrington, " The Integrative Action of the Nervous System ": New York 1906, Cambridge 1947.)

pyramidal tracts descend in the internal capsule between thalamus and basal ganglia ; they occupy part of the cerebral peduncles on the ventral surface of the midbrain, pass through the pons and emerge to form two swellings, the pyramids, on the ventral surface of the medulla. Most of the fibres cross the midline in the pyramidal decussation to enter the opposite lateral column of the spinal cord (Fig. 14. 12). A fraction (of unknown function) remain uncrossed and pass down the cord on the same side, mainly in the ventral column. Other fibres initially taking a similar course branch away to the motor nuclei of the cranial nerves.

The largest and fastest conducting fibres arise from exceptionally large neurones called Betz cells, of a size not found elsewhere in the cortex, but these are in a minority ; by far the greater number of fibres from the motor area arise from ordinary small cortical cells. The pyramidal tracts (defined as the tracts that form the medullary pyramids) also contain very many corticospinal fibres (of obscure function) from the so-called premotor areas in front of the motor areas and from the sensory areas and adjacent parts of the parietal lobes behind them. In all there are more than half a million fibres in each pyramidal tract in man, but only some 30,000 Betz cells in each motor area.

The motor area was historically the first part of the cerebral cortex that was clearly shown to have a specific function. As so often in the central nervous system the first evidence came from the study of human patients. The central nervous system in its bony case is so difficult for the experimenter to get at and the results of operating on it are so difficult to evaluate in animals, who cannot be given instructions to attempt various movements or be asked to describe what sensations they have, that in many instances the first clues to the function of a part of the brain have come from the destructive experiments that disease performs in man. In fact to this day the parts of the brain about which we are most ignorant, such as the basal ganglia of the cerebral hemispheres or the olives in the medulla (see p. 436 and Fig. 14. 16), are mainly those which do not suffer localised clear-cut destruction in disease.

With the cerebral cortex it is not only destructive lesions that provide evidence, for the cortex is also prone to those paroxysmal bursts of overactivity which are the cause of epileptic fits (convulsions). This sort of instability seems to be peculiar to the cortical grey matter. The unsolved question of what happens in the cortex during an epileptic fit is a matter of physiology, not pathology, for anyone will have a fit if his brain is stimulated sufficiently violently. A convulsion triggered off by passing 50 cycles per second alternating current through the head is a standard and effective method of treatment in various psychiatric disorders, while in normal infants convulsions not uncommonly occur during teething, or ushering in an acute infective illness. In sufferers from epilepsy similar convulsions occur spontaneously. In a typical fit the whole cortex is involved more or less simultaneously, the patient loses consciousness and then has a generalised convulsion. Much less often convulsive movements begin in one part of the body only and spread to involve the remainder of that side of the body and then the other side.

Epilepsy is usually not associated with any changes in the naked-eye or microscopical appearance of the brain, but sometimes it is precipitated by obvious damage, a cerebral tumour or the scars of a brain injury. In the early 1860s Hughlings Jackson observed that convulsions starting unilaterally (still known as Jacksonian fits) were associated with damage to the precentral region of the opposite hemi-

sphere. When the fit began in the foot the lesion was at the top of the precentral region ; with lesions in the middle the hand was involved first and so on. The commonest sites for the fits to start were the feet, the fingers and hand, and the corner of the mouth. From these facts Jackson inferred that the precentral gyrus is the cortical centre for movements, and that within it the body is represented upside down with the largest areas concerned with movements of the important " leading parts." The manner of spread of the convulsion, the " march" as he called it, reflected the spread of epileptic activity over the motor cortex, involving one area after another.

Soon after this Fritsch and Hitzig showed that movements of the opposite limbs could readily be obtained by electrical stimulation of the precentral cortex in the dog. Previous to this it had been generally held that the cerebral cortex was electrically inexcitable. Before the end of the century electrical stimulation had been used to map out the motor area in many species of animals and in man.

The paralysis caused by a lesion of the motor cortex is of a highly distinctive character and distribution. The actions lost or most severely weakened are the most voluntary and least automatic, the most accurately graded and individual movements in which the limb is usually employed, the sort of movements out of which skilled actions such as writing or driving a car are made up. Movements of this type employ predominantly the distal parts of the limbs, particularly the hands and it is here that the paralysis makes itself most felt, the proximal parts are less affected, while bilateral movements of the trunk are preserved. As already mentioned (p. 394), expressive and emotional movements of the face tend to be much less affected than movements made to order.

Within the hand itself the same principle applies ; the most highly specialised movements are the most paralysed. The ability to move single fingers independently of the others is the first thing to go, so that the patient may lose manual dexterity completely while retaining a powerful grip. The same muscles move each finger whether they are flexed one at a time or all together, but the one movement is lost and the other is not. From this and many similar instances it is clear that it is not the muscles themselves that are paralysed but only certain movements carried out by them. The motor cortex, even more obviously than the spinal cord, " thinks " in terms of movements, not muscles.

Electrical stimulation of the human motor cortex exposed at operation under local anæsthesia causes fragmentary movements of just the type that are lost when the same area is removed. They predominantly involve the distal parts of the limbs, fingers and toes, wrist and ankle, and include a large proportion of discrete movements of the thumb and individual fingers of just the sort that a patient with a cortical palsy cannot make. Similar movements can be obtained after the grey matter has been removed, by stimulating the cut ends of the corticospinal fibres in the white matter underneath ; hence stimulation of the cortex reveals mainly the properties of the corticospinal pathway and

if, as seems probable, the cortex itself has the function of organising complete actions from combinations and sequences of these fragmentary movements, this activity is not put into motion by the type of electrical stimulation employed.

Some writers have argued that movements are represented in the motor cortex in the same sense that movements could be said to be represented in the skin ; appropriate patterns of excitation can call forth elaborate muscular actions, but simple electrical stimulation of the skin or the cortex only discovers the most elementary. The analogy is a good one, for corticospinal fibres mainly end on interneurones and not on the motoneurones themselves. Both appear to activate the spinal mechanisms that synthesise functional movements out of contractions of individual muscles. The ease with which movements of the digits are obtained by cortical stimulation is partly a property of the spinal cord, for these are also the movements most readily obtained in spinal men and spinal monkeys by mechanical stimulation of the skin (*e.g.* the extensor plantar response, p. 405) or by electrical excitation of the sensory nerve fibres. Thus the great importance of the distal parts of the limbs in voluntary life is reflected in the spinal motor mechanisms, and it is the large area of cortex connected to these mechanisms that cortical stimulation reveals.

In man voluntary paralysis is commonly the result of rupture of fibres leaving the motor cortex by hæmorrhage from a burst blood vessel near the internal capsule. The patient has an " apoplexy " or " stroke," *i.e.* suddenly has his senses and his power of motion taken from him, and is subsequently found to be paralysed down one side of his body (a hemiplegia). The plantar response (p. 405) on the paralysed side becomes extensor (upgoing big toe) from the onset. As usual after any sudden damage to the brain a good deal of recovery of function takes place in the next few weeks, possibly due to recovery of structures that were put out of action but not actually killed. Initially the paralysed limbs are quite flaccid, but if voluntary movements do not recover they gradually become stiff (spastic) owing to exaggeration of the stretch reflexes (p. 407), and the tendon jerks (p. 411) are increased. There is evidence that the spasticity is not due to destruction of corticospinal fibres ; thus it is said that following division of the corticospinal tract in the medullary pyramid in a monkey, the paralysed limbs remain flaccid. With a lesion in the internal capsule many fibres other than corticospinal fibres are divided. The important ones for spasticity may be those from the motor and premotor areas to motor nuclei in the brain stem belonging to the extrapyramidal system (p. 436 and Fig. 14. 16). Why division of such fibres should cause spasticity is unknown. Spastic paralysis with increased tendon jerks also results from damage to the spinal cord short of complete division (paralysis in both legs due to a spinal lesion is called paraplegia). How this type of spasticity is related to that due to disease in the brain is not known. In view of these uncertainties, it is safest not to speak of spasticity and increased jerks, when they occur with lesions of the brain or spinal cord, as indicative

of damage to the pyramidal tract, but to use some non-committal term such as " long motor tracts."

Another expression used is " upper motor neurone," dating from the days when the voluntary motor pathway was considered to consist of two types of neurones only, the upper motor neurone (the Betz cell in the motor cortex with its axon) connected to the lower motor neurone (the motor cell in the ventral horn of the spinal grey matter with its motor axon running to a muscle). Spastic paralysis with increased tendon jerks and an extensor plantar response are spoken of as signs of an upper motor neurone lesion. Lesions of the lower motor neurone (*e.g.* poliomyelitis) cause a flaccid paralysis with loss of tendon jerks and eventual wasting of the denervated muscles. There is no wasting in an upper motor neurone lesion.

As pointed out by Hughlings Jackson, destruction of nervous tissue can only be the direct cause of loss of function, of purely negative symptoms such as paralysis. Positive symptoms such as spasticity and the change in character of the plantar response must be due to the action of surviving parts of the nervous system that are " released " from influences that previously controlled them. Release phenomena appear to fall into two categories, those that come on immediately after a lesion (*e.g.* the extensor plantar response after a stroke, or decerebrate rigidity after section of the midbrain in an animal), and those that develop after an interval of weeks or more (*e.g.* spasticity, and the thalamic syndrome (p. 427)). The hypersensitivity of denervated muscle to acetylcholine (p. 384) may also be considered as a release phenomenon of the second class. The relationship of clinical spasticity to the very similar state of the muscles in decerebrate rigidity in animals is not understood.

Speech. Speech is a form of motor activity peculiar to man. Inability to speak which is not caused by muscular paralysis or other local disease of the larynx, tongue, etc., is called aphasia. Patients with aphasia are often unable to understand what is said to them, or to read and write or to do arithmetic. These related disabilities are conventionally dealt with under the heading of aphasia. In right-handed persons aphasia is caused by lesions of the left cerebral hemisphere. Damage to the cortex in front of the motor area for the face at the bottom of the precentral gyrus (Broca's area, Fig. 14. 3) causes a loss of speech with no loss of comprehension of other people's speech or of writing, a so-called motor aphasia. Damage to the parietal lobe in the region between the receiving areas for vision and hearing gives rise to a receptive aphasia ; the patient cannot understand speech and may not be able to read. Words mean nothing to him and he cannot speak because he has no words to express his thoughts. In this type of aphasia he may talk jargon and fail to recognise that what he is saying is meaningless. Although difficult to assess in an aphasic it is thought that intelligence generally suffers. Thus aphasia may be only one symptom of a general inability to use symbols in thinking.

One of the many interesting things about aphasia is that the use of

words to express emotion is often preserved. The patient who cannot get out the simplest sentence to express a thought, may yet be able to swear or to express approbation and disapprobation with a simple yes and no. We have already noted that movements of the face to express emotion may likewise be preserved after damage to the cortical motor pathways.

The Frontal Lobes. The frontal lobes are the parts of the cerebral hemispheres extending forwards of the central sulcus. They therefore include the motor and premotor areas already discussed. Extensive damage or removal of one or both frontal lobes ahead of the motor and premotor areas does not cause any immediately obvious disability. Such a patient displays no defect of sensation, no paralysis of movement or loss of skill and no marked impairment of memory or intellect. Further observation, however, reveals that his personality has changed, the changes being much more conspicuous if both frontal lobes are involved. He is no longer able to plan ahead or to pursue any but the most immediate goals. Even if cautious and far-sighted before the operation he is now happy-go-lucky. His insensibility to the consequences of his acts may be embarrassingly apparent in his social behaviour. He makes jokes in bad taste, quarrels readily and may pass his water into his trousers, all with a blithe unawareness of the awkwardness he causes. These facts suggest that the frontal lobes are concerned with the long term planning of behavioural activity, with appreciating the consequences of various courses of action and with the emotional concomitants of this sort of activity, such as anxiety.

The operation of prefrontal leucotomy, cutting the white matter which contains the fibre tracts connecting the frontal lobes with the rest of the brain, may be of benefit to certain types of mentally sick patient, particularly those incapacitated by obsessions and by pathological anxiety and suspicion. The general tendency of leucotomy is to make the patient carefree (see also p. 448).

Subcortical Motor Mechanisms

It is easy to gain the impression that there is little more to voluntary movement than the discharge of motor impulses down the pyramidal tract to the spinal motor mechanisms. The very large contribution that subsidiary mechanisms make to ensuring successful execution of the desired act tends to be lost sight of because, like the perfect servant, we are not conscious of it. The parts of the brain involved are, firstly the basal ganglia and neighbouring nuclei in the midbrain, known collectively as the extrapyramidal system, and secondly the cerebellum with the structures in the brain stem related to it (Fig. 14. 16). In addition there are automatic mechanisms based on spinal reflexes which appear to regulate the contraction of individual muscles in all types of activity, reflex and postural as well as voluntary.

The Extrapyramidal System. There are wide differences of usage in the term extrapyramidal. The cerebellum is strictly extrapyramidal, but one common convention which is adopted here treats it separately

and confines the term extrapyramidal to the basal ganglia (the corpus striatum) and the midbrain nuclei closely connected with the basal ganglia, notably the sub-thalamic nucleus, the substantia nigra and the red nucleus, to which should probably be added some of the cells in the reticular formation of the brain stem (see also p. 444). In lower vertebrates in which the cerebral cortex is little developed the basal ganglia are the highest motor centres and the rubrospinal tract from the red nucleus is the principal tract to the spinal cord. As the cortex and pyramidal tracts increase in size the rubrospinal tract becomes less important and in man it is small. At the same time the amount an animal can do after the pyramidal tracts are sectioned in the medulla decreases. A cat moves about and performs many acts normally, but movements of individual limbs, pawing objects, etc., are lost. The monkey loses all delicate limb movements but after some weeks it can stand, and move about unwillingly. Man, no doubt, could manage even less.

These activities are little more than might from analogy with the cat and dog be attributed to action of the midbrain alone and scarcely seem to require the large development of the basal ganglia in primates. The evidence of disease in man suggests that the basal ganglia, having resigned the leading *rôle* in the motor system, have been taken into partnership by the cortex ; but exactly what they contribute is very obscure. Disease of the extrapyramidal system tends unfortunately to be diffuse and to affect unevenly many of the nuclei partially destroying them, but it is clear at any rate that unilateral disease of the basal ganglia results in symptoms on the opposite side of the body (the rubrospinal tracts decussate in the midbrain).

The commonest affliction of the extrapyramidal system is paralysis agitans (Parkinson's disease) in which the limbs become stiffened at all joints by contraction of both flexors and extensors and develop tremor at a frequency of three or four oscillations per second. The muscular contractions involve the stretch reflex but tendon jerks are not increased and the clasp-knife response is not obtained. This type of stiffness is called *rigidity* to distinguish it from the *spasticity* ensuing on damage of the long motor tracts (p. 434). What a physiologist calls decerebrate rigidity would be decerebrate spasticity to a clinical neurologist.

In Parkinson's disease the patient is disabled because he cannot get his voluntary movements going ; all movements become cramped and slow and he walks with tiny shuffling steps. In the early stages the first things noticed even before rigidity and tremor develop are slowness in manual tasks, such as getting dressed, a dead-pan expressionless face, and a failure to swing the arms when walking unless attention is given to them. It appears that many automatic elements in movement are lost and are slowly replaced by rigidity and tremor. Rigidity and tremor, in Hughlings Jackson's terminology, are release symptoms and they must be due to overaction of surviving nervous tissue. What is overacting in Parkinson's disease may perhaps be the surviving parts

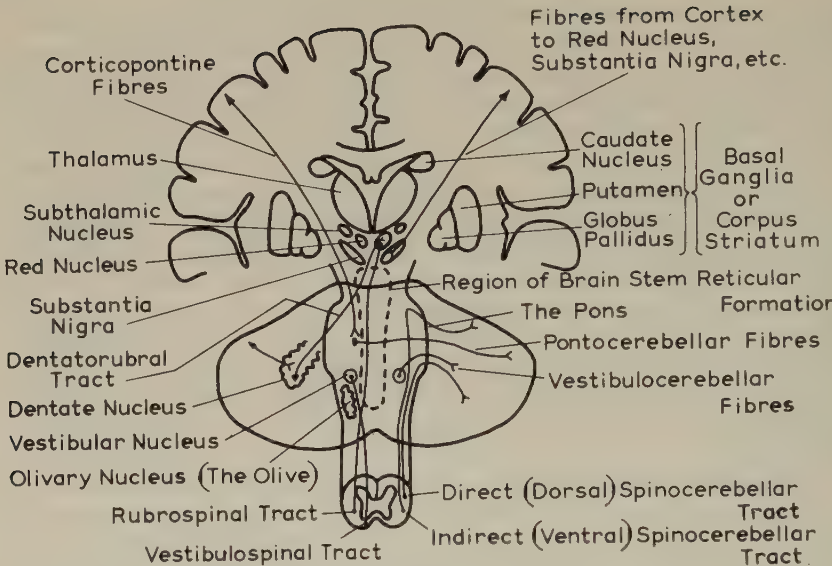


FIG. 14. 16. The extrapyramidal system and the cerebellum. The nuclei are shown on a diagrammatic section of the human brain in the plane of the brain stem as in Fig. 14. 11, with, here, the outline of the cerebellum and one dentate nucleus superimposed. Some structures not labelled here are labelled on Fig. 14. 11 ; for spinal tracts see also Fig. 14. 12.

The extrapyramidal nuclei are connected to each other and to the thalamus by numerous bundles of nerve fibres (not shown). The main tracts running to them are the dentatorubral from the cerebellum and fibres from the cerebral cortex to the red nucleus and substantia nigra. There appear to be relatively few direct connections between the cerebral cortex and the corpus striatum. Likewise the efferent connections of the extrapyramidal nuclei to the spinal cord arise mainly from the red nucleus and reticular formation. The reticulospinal tract (not shown) runs near the rubrospinal tract.

The olives are large nuclei of unknown function (only the left olive is shown), which receive large tracts of fibres from the neighbourhood of the red nuclei. They have efferent connections to the opposite cerebellar cortex and to the spinal cord by the olivospinal tracts (not shown), which run near the vestibulospinal tracts.

The cerebellum lies dorsal to the hindbrain connected to it by three peduncles (not distinguished in the diagram) on each side : superior, containing mainly the large dentatorubral tract and the indirect spinocerebellar tract ; middle, the largest in man, consisting of pontocerebellar fibres from the pons ; and inferior, containing principally the direct spinocerebellar tract, vestibulocerebellar fibres, and (not shown) fibres from the dorsal column nuclei to the cerebellar cortex, and efferent fibres (not shown) from the central nuclei other than the dentate to the medulla.

of the basal ganglia, for further destruction of one of the basal ganglia, the globus pallidus (Fig. 14. 16), by injecting alcohol into it, has been found on occasion to relieve the rigidity.

There are facts which suggest that a patient with Parkinson's disease is not so much incapable of moving as prevented from doing so. It is of great interest that some gravely disabled cases may, if suddenly

and powerfully motivated, move swiftly and effectively. Thus a patient who has great difficulty in feeding himself or walking may deftly catch a ball thrown to him, or jump out of the way of a car about to run him down. Gordon Holmes saw " a man who could scarcely walk in his waking state wander about easily in a period of somnambulism." Such dramatic temporary recovery would be unthinkable in patients with damage to the pyramidal pathways or the cerebellum.

In other patients disease of the corpus striatum results in involuntary movements of the limbs, which may take the form of elaborate writhing movements (athetosis) or of irregular jerkings (chorea), which may be of sufficient violence to damage the limbs. Which precise structures have to be destroyed to produce the rigidity-tremor syndrome and which the involuntary movements of athetosis or chorea is not known. This reflects the generally unsatisfactory and confused state of present knowledge of the basal ganglia.

The Cerebellum. The cerebellum appears early in evolution as a piece of private brain belonging to the vestibular-lateral line system of sense organs, which are believed to be important in the control of swimming. In primitive animals the cerebellum consists of two separate outgrowths, one on each side of the medulla, each connected to the vestibular nerve on that side. In higher animals they enlarge and fuse in the midline, but the connections to the cerebellum from sense-organs remain predominantly uncrossed, so that each half of the cerebellum deals with the half of the body on the same side.

In higher animals the cerebellum has become the civil service of the motor system ; having started with the job of seeing that swimming was properly carried out having regard to all the circumstances communicated to it by the vestibular-lateral line system, it later took over the executive side of voluntary movements as well and uses a wider range of sense data to see that they are carried out properly too.

The cerebellum is a large organ in animals that conduct their lives in three dimensions, *e.g.* fish, birds and whales, as opposed to those operating essentially on a surface, *e.g.* frogs. But it reaches its greatest size in the higher mammals with their elaborate voluntary motor activities. In them it is not the vestibular parts of the cerebellum that have developed but mainly phylogenetically newer portions, the enlargement of which parallels the great development of the cerebral hemispheres. In man the vestibular parts are completely overshadowed.

In mammals the cerebellum receives afferent impulses from the vestibular organs, from the limbs and trunk (via the direct and indirect spinocerebellar tracts (Fig. 14. 16) and to a lesser extent via the dorsal columns), and from the eyes and ears. The visual and auditory connexions are via the superior and inferior colliculi respectively. Information about what the motor cortex is doing is sent by very numerous fibres that run near the corticospinal tract as far as the pontine nuclei whence fresh fibres forming the pons and the large middle cerebellar peduncle are distributed to the opposite cerebellar cortex. The cere-

bellar cortex consists of a uniform thin sheet of grey matter the area of which is enormously increased by deep and elaborate folding. The efferent fibres do not arise directly from the cortex but from central masses of grey matter, of which the largest are the dentate nuclei, one on each side. The principal efferent tracts run in the superior peduncles to the red nuclei ; others go via the inferior peduncles to motor centres in the reticular formation in the medulla.

In the cat and monkey afferent impulses from the hind limbs go to the anterior part of the cerebellum (lobulus centralis and front part of culmen), from the fore limbs to the part behind this (culmen and

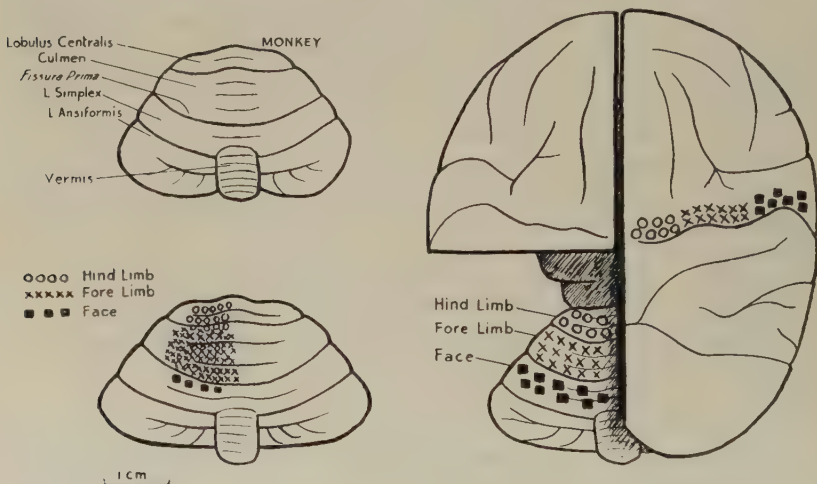


Fig. 14. 17. Regions of the monkey's cerebellum connected with the limbs and face, and with the motor cortex.

Top left, the anatomical divisions of the dorsal surface of the monkey's cerebellum.

Bottom left, the regions in which impulses can be detected by electrical recording when various parts of the body are moved or pressed upon.

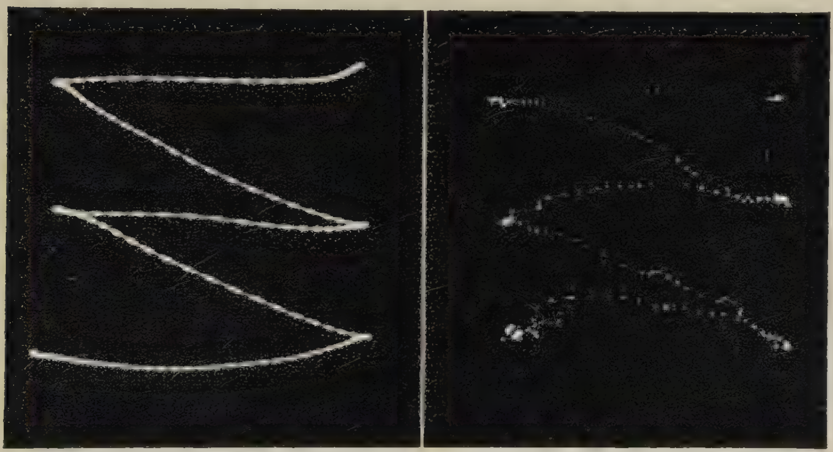
Right, the regions in which discharges originating in the limb and face areas of the motor cortex can be detected arriving in the cerebellum via the cortico-ponto-cerebellar fibres.

(Adrian, *Brain*, vol. 66, p. 289, 1943.)

sometimes part of lobulus simplex) and from the face to the part behind this again (lobulus simplex). Fig. 14. 17 shows these areas in the monkey. The pathways to the cerebellum are uncrossed, the left fore-limb sends impulses to the left half of the culmen, and so on. Discharges to the limb areas are most easily produced by movement of joints and muscles and by pressure on the pads of the feet, but light touch is sometimes effective. The arrival of impulses is detected in the same way as in the sensory area of the cerebral cortex (p. 428). As with the cerebral cortex the pattern of impulses that arrives is very much like what it is in a peripheral nerve fibre. Auditory and visual stimuli cause activity behind the face area.

The cortico-ponto-cerebellar pathway as a whole projects to a larger area than the afferent pathways just described, but within this area the hind-limb, fore-limb and face areas of the motor cortex are connected to the cerebellar areas that receive afferent impulses from the same parts of the body (Fig. 14. 17). It is worth noting that the mapping of the body on to the cerebellar surface continues across the morphologically important division between anterior and middle lobes, which occurs between culmen and lobulus simplex.

The cerebellum is wholly concerned with the execution of muscular acts ; it has no say in what is to be done, policy making is for the cere-



A B

FIG. 14. 18. Records to illustrate cerebellar ataxy. A flashing light was attached to the tip of a forefinger and its movements photographed in a dark room by leaving the camera shutter open. The light flashed 25 times per second ; thus the record shows not only the track of the forefinger but also, by the spacing of the flashes, its velocity.

The task was to move the forefinger accurately between two columns of illuminated points (which do not show up in the photograph), some 75 cm. apart. The patient had right-sided cerebellar ataxy. A is the record from the unaffected forefinger, B that from the right forefinger. Record B shows intention tremor as each target point is approached, and irregularities of rate and direction in between. (Gordon Holmes, *Brain*, vol. 62, p. 1, 1939.)

bral cortex. The results of damage to the cerebellum are disturbances of posture and disorganisation (ataxy) of movement. If the lesion is on one side of the cerebellum the symptoms are on the same side of the body. In man nothing that goes on in the cerebellum reaches consciousness and damage to the cerebellum does not affect memory, intelligence, character, or any other aspect of our mental life. Although it is richly supplied with afferent information, damage to the cerebellum causes no sensory loss. Neither is the patient paralysed, he is not unable to initiate movements, but the movements he begins do not come off properly. At rest little or nothing is observed to be wrong. The characteristic disturbances appear when the patient is asked to

perform some action, *e.g.* to touch an object in front of him with his index finger. He starts off without difficulty but his hand soon deviates from the correct line and may change direction and velocity several times in an irregular jerky manner. His finger may miss the target altogether, often by overshooting it, and wobbles irregularly as it approaches or even after it has made contact, a symptom called intention tremor (Fig. 14. 18). The patient knows that his arm is moving incorrectly and may complain spontaneously that he is unable to control it ; he tries to move it in one direction, he says, and it goes off in another. As already implied this is not due to loss of sensation in the arm ; in particular it is not due to loss of position sense ; the patient knows where his arm is well enough and he does not have to look at it to find out. Cerebellar ataxy is not aggravated by shutting the eyes, unlike the superficially similar sensory ataxy caused by lesions of the dorsal columns and elsewhere.

Examination shows that the muscles in the ataxic limb are contracting and relaxing at the wrong moments and exerting the wrong force when they do contract (dysmetria). Much of the trouble is due to failure of the muscles to steady the shoulder girdle and the elbow ; the arm swings like a barn door in the wind and the intention tremor is partly due to the patient's voluntary attempts to compensate for this lack of the normal postural fixation of the proximal joints. At rest the arm offers no resistance to passive movement ; stretch reflexes are absent although the tendon jerks are normal (p. 411). These observations show that the cerebellum has the task of making the automatic modifications of posture that we take for granted when we make voluntary movements, and for arranging that the numerous muscles involved in a movement contract at the right time and with the right force. One possible way in which the cerebellum supervises muscular contractions is discussed on pp. 444–5. The cerebellum also ensures that successive components of a complex act run smoothly into each other ; in cerebellar ataxy there is particular difficulty with alternating movements, delays occur at each turning point ; similarly, in speaking, words are broken up and each syllable pronounced separately (scanning speech).

Nervous Control of Muscular Contraction. Until recently it was often thought that orders from the cortex to carry out a voluntary movement were delivered by pyramidal fibres direct to the motoneurones of the muscles concerned, and that even in reflex action impulses in sensory nerves made their way straight to the motoneurones. We have seen that in general this is not so ; most pyramidal and dorsal root fibres end elsewhere in the spinal grey matter and set in motion spinal mechanisms that elaborate the motor act. Almost nothing is known of how complex acts such as scratching and walking are produced, but evidence has been cited (pp. 407 and 417) that they are not in the nature of chain reflexes ; the basic mechanisms seem to be built into the grey matter and can function after a fashion without impulses from the peripheral sense organs.

One common feature that stands out is that when a limb is moved the muscles in it that would oppose the movement are caused to relax. This is known as *reciprocal innervation*; it is seen very clearly if a flexion reflex is elicited in the leg of a decerebrate cat. The rigid extensor muscles of the knee immediately relax; this occurs even if the flexors of the knee are detached from their insertion so that the knee is not moved, or even if the motor nerves to the flexors are cut so that they do not contract at all. The mechanism of reciprocal innervation appears to be in the spinal grey matter, but it is assisted as usual by reflexes from the periphery. Thus intracellular recording from motoneurones (p. 380) has shown that impulses from muscle spindles in flexor muscles inhibit extensor motoneurones. Movements elicited by electrical stimulation of the motor cortex also exhibit reciprocal innervation.

Reciprocal innervation was of great importance in the development of neurophysiology because it revealed to Sherrington the large part played by inhibitory processes in nervous action. Not all types of muscular action, however, employ reciprocal innervation. It is very clear that flexors and extensors can readily be made to contract together by voluntary effort, and the same thing occurs automatically to fix the joints in standing (p. 416). In these instances the limbs are stationary; reciprocal innervation comes into play when they move.

Even the spinal mechanisms that arrange for the correct muscles to contract in the correct sequence may not act merely by exciting their motoneurones: there is evidence which suggests that in many instances they activate subsidiary mechanisms whose job it is to regulate the contractions of the individual muscles. As already mentioned on p. 409, muscles engaged in all kinds of contractions, reflex, postural and voluntary, are found to show the stretch reflex. There is evidence that when a muscle shortens impulses from the γ motoneurones make the muscle spindles shorten too, so that at whatever length the muscle reaches the stretch reflex remains active (see Fig. 14. 8). In fact, in reflex contraction, direct recording from single nerve fibres has shown that the γ discharge begins before the α motoneurones start up. The first thing that happens, therefore, is that the spindles contract, which stretches the sensory region and sends impulses up the monosynaptic afferents. Such impulses are excitatory to the motoneurones and assist in causing them to discharge. It would be possible, of course, for the motoneurones to be caused to discharge by thus turning on the stretch reflex without any excitation reaching them from elsewhere, but whether contractions are wholly driven via the γ route in this way is not known. Probably not, some direct excitation of α motoneurones appears likely; the most that can be said is that in reflex activity in the decerebrate cat the γ route contributes a substantial fraction.

To an engineer the ordinary stretch reflex is a servo-mechanism tending to hold the muscle at a certain demanded length by using negative feedback from change-of-length detectors in the muscle. If the load varies it turns muscle tension on or off in an attempt to maintain the demanded length. This is very clearly seen in an extensor muscle

of a decerebrate cat. A shortening of a muscle driven or assisted by spindle contraction is equivalent to changing the demanded length. Thus we arrive at the idea that this servo-mechanism based on the stretch reflex allows the motor centres to order a certain change of length, which furthermore will be carried out with the same automatic adjustment of tension to load that is observed in the ordinary stretch reflex. In many actions what is required is a change of limb position, *i.e.* changes of muscle length. A muscle, on the other hand, is essentially a device for producing tension ; feedback from the muscle spindles on to the motoneurones appears to modify the characteristics of the system so that when excited via the γ route it tends instead to produce changes of length, by servo action.

There is certainly a great deal that is not yet discovered about the functions of muscle sense organs, and how they are employed to make the muscles do what is required of them. The above account will no doubt prove to be at least incomplete in many directions ; but it ought to make clear why the stretch reflex is coming to be thought of as part of the muscles's own nervous machinery ; and why, when this machinery depends on excitation of the γ motoneurones, in addition to the α motoneurones, it is missing the point (if not exactly incorrect) to think of all reflex and voluntary pathways as converging on to the " final common path " (p. 404) of the α motoneurones.

Not all muscular contractions are based on the stretch reflex. It is well known that many spinal actions can be obtained after cutting all the dorsal roots to the moving limb, which necessarily abolishes all stretch reflexes in it. Examples already given are the crossed extensor reflex (p. 403), the scratch reflex (p. 407) and spinal walking (p. 417). Voluntary movements can also be carried out by a deafferented limb. In these instances all the excitation must be reaching the α motoneurones direct. Granit and his collaborators have now shown that even with the dorsal roots intact the nervous system has the choice of employing this direct α route instead of turning on the stretch reflex by exciting the γ motoneurones. Which route is employed appears to depend on the cerebellum. The evidence for this derives from experiments on decerebrate rigidity, which must now be discussed.

Decerebrate rigidity develops in the cat, dog, etc., when the midbrain is sectioned behind the superior colliculus. As described by Sherrington in 1898, such rigidity disappears in a limb if the dorsal roots to that limb are cut, because, as we now know, it depends on exaggerated stretch reflexes. Much later it was observed that, if the anterior lobe of the cerebellum was removed too, an outwardly similar rigidity developed, but it was not abolished by section of the dorsal roots. It, therefore, could not be due to exaggerated stretch reflexes. The mechanisms in the brain stem and cerebellum responsible for these phenomena are not understood. (Magoun and Moruzzi and their colleagues have investigated the excitatory and inhibitory effects on spinal reflexes which can be obtained by stimulation of two regions in the reticular formation. An explanation of decerebrate rigidity is

sought in terms of a disturbance in the balance of such antagonistic reticular influences.) At all events it is clear that there are two distinct kinds of decerebrate rigidity, and it turns out there are two kinds of reflex movements also. Single fibre recording has shown, as already described, that during reflex movements vigorous contractions of the muscle spindles occur in the ordinary decerebrate animal; after removing the anterior lobe of the cerebellum practically indistinguishable reflexes can be elicited but the spindles remain passive. Thus the reflex centres may activate the muscles in either of two ways. Probably varying combinations of the two routes are normally employed. There is some evidence that the same is true of voluntary contractions in the human.

It is a matter of conjecture what the advantages of this double system are to an animal. The direct route may be better in rapid movements because of the extra delays involved in conduction time with the γ route. The desirable properties that the stretch reflex confers on a muscle have been emphasised already. It has been suggested that the dysmetria and failure of joint fixation in human cerebellar ataxy may be due to a loss of the type of contraction driven through the γ motoneurones, to spindle paralysis in fact. One of the ways in which the cerebellum supervises muscular activity may be by making sure that the best route, the α route or γ route, is used to activate the muscle. At all events it seems justifiable to consider the details of muscular contraction separately from the question of which muscles contract. The local spinal mechanisms belonging to the muscles and the cerebellar mechanisms that control them appear to be at the disposal of both reflex and voluntary, and probably of postural mechanisms as well.

Learning and Memory

A man's present actions may be influenced by what happened around him, by what he thought and felt, and by what he learnt to do many years ago. The question of how and where the brain stores memory traces of these things is of great general interest, but so far only hints of the answer are available. As regards how, no electrical phenomena, action potentials, changes in synaptic excitability, etc., are known which last longer than a few seconds. It has therefore been suggested that memories depend on nerve impulses continually circulating over closed pathways, the nature of the memory being determined by the pattern of the pathway. This theory runs into the difficulty that memories cheerfully survive events such as profound coma or epileptic fits which would be expected to stop or disorganise the electrical activity of the brain. For these reasons it seems probable that long-lasting structural changes are involved, using the term in the widest sense to include such things as changes in the amount of an enzyme in synaptic terminals. The possible nature of these changes is a matter of speculation only at the present.

A fact that any theory has to account for is that recent memories

are more easily lost than old ones. It is a commonplace that old people may recall endless details of their childhood after they can no longer remember the events of middle age, or even of a week ago. Similarly after a head injury with concussion the patient may remember nothing of the events leading up to the accident ; he has what is called a retrograde amnesia ; commonly it goes back a few minutes or an hour or two but in severe cases it may extend to several weeks before the accident. Retrograde amnesia is one instance of a general principle, first clearly stated by Hughlings Jackson, that the most recently acquired and the most highly organised activities are the first to be lost when the brain is damaged. A hemiplegic loses the ability to write while he can still pick up his bread and butter. The extensor plantar response (p. 405) he exhibits represents a return to the infantile form of the reflex. The emigré who develops a degree of aphasia reverts to his mother tongue, which he may not have spoken for years.

Loss of memory is an early symptom of generalised shrinkage of the cortical grey matter when it occurs in old age or in various forms of dementia. The demented patient forgets past events, names, faces, words, the time, where he is, what he was going to do next, and so on. He also loses his code of social behaviour, his power of reasoning and his ability to cope with new situations. One might say he has forgotten how to behave, how to think and how to adjust. Many of these disabilities are presumably due to loss of the learned activities of some of the special areas of the cerebral cortex already described. A patient with damage to the frontal lobes can be said to have forgotten how to behave. The aphasic cannot recall names or words. After damage to the parietal lobe a patient may forget where he is. Regional cortical lesions of this kind, however, seem never to cause a loss of memory for past events. It is therefore sometimes suggested that memory for past events is a function of the cortex as a whole.

Support for this view came from Lashley's experiments with rats in the nineteen-twenties on the learning and retention of simple habits, such as threading a maze to obtain food. He excised various parts of the cerebral cortex and came to the conclusion that it did not matter which part he removed ; the impairment in learning appeared to depend only on the total area of cortex removed. The only sign of localisation was that tasks involving visual pattern recognition required the visual areas. Lashley's results could be explained if the learning process involved every part of the cortex.

In man the memory traces of past events may be equally widely spread out, but against this is a certain amount of evidence that the temporal lobes are specially involved in memory of this kind. In epilepsy beginning in the under part of the temporal lobe (uncinate fits) the patient may experience feelings that what is going on around him is intensely familiar ; it has all happened before (déjà vu). Or he may vividly relive incidents in his past life. Recently it has been found that if the anterior parts of both temporal lobes are removed (as was done in an attempt to cure epilepsy) new memories cannot be

formed. The patient cannot remember what happened even a few seconds ago and lives literally for the moment. Intellect, sensation, and other cortical activities appear unaffected.

Conditioned Reflexes. Reflexes such as the knee jerk, or constriction of the pupil when light is shown in the eye, are innate, but many other automatic actions are learnt. If a dog is given food and at the same time a bell is rung and this is repeated several times at intervals of, say, five minutes, it is found that after some time the dog will salivate when the bell is rung, without any food being offered. The dog's nervous system has learnt to associate the bell with food and a new reflex is established. The famous Russian physiologist, Pavlov, called the process of acquiring such reflexes " conditioning " and the reflexes " conditioned reflexes." Innate reflexes were called " unconditioned." Most reflexes that depend on visual and auditory recognition by the cerebral cortex are probably conditioned, not innate. Salivation when food is placed in the mouth is an unconditioned reflex. In addition all normal dogs, if hungry, will salivate when merely shown meat. Pavlov showed that this is a conditioned reflex, for if dogs are reared without ever eating meat they do not salivate when shown it for the first time, although, if allowed to eat it, they salivate at once. In man such reflexes as the blink to a threatening movement, and an enormous number of other automatic responses to events signalled by the eyes and the ears, are doubtless conditioned reflexes established in childhood. Only the very simplest conditioned reflexes remain or can be established after removal of the cerebral cortex, but innate reflexes are not affected.

Since the beginning of the century Pavlov and his school have intensively studied conditioned reflexes, working chiefly with reflexes causing salivation in the dog. One parotid duct was transplanted to the skin of the cheek so that the saliva could be collected, and its amount measured. Thus the experiments were made roughly quantitative. Conditioned reflexes are very easily prevented by any disturbance. Thus, if someone walks into the experimental room, or if a buzzer is sounded, the conditioned secretion of saliva to ringing of a bell is likely to fail on that occasion. For this, and other reasons, the dogs were put in soundproof rooms with the experimenter invisible outside, and all manipulations, etc., were done by remote control. The prevention of conditioned reflexes by miscellaneous interruptions is called " external inhibition." The other sort, " internal inhibition," occurs, for example, if (having established a conditioned reflex) the bell is rung at intervals but food never follows. After some repetitions the flow of saliva dwindles and ceases. It would not be very surprising if the association were lost under these circumstances, but there is good evidence that the reflex is not so much lost as actively suppressed. For instance, it recovers very rapidly if food is presented with the bell once or twice. This is called " reinforcement." It also recovers spontaneously if the dog is left for some hours. But most striking of all, it reappears at once if a stimulus, such as an electric buzzer, of the kind that usually causes external inhibition, is given at the same time as the bell. It seems that

the first thing external inhibition acts on is the last thing the animal learnt, namely the association of the bell with no food which overrode the original reflex. Hence the conclusion that the original reflex was inhibited, not lost.

Animals can be trained to make sensory discriminations in their conditioned reflexes. For example, if a musical note of a certain pitch is established as a conditioned stimulus, then initially many other notes will also excite the reflex. But if the first note alone is reinforced (*i.e.* followed by food) while the others never are, sooner or later only the original note is effective. Pavlov thought of this as an internal inhibition to the other notes. When an indifferent disturbing stimulus is given, such as the buzzer, discrimination tends to be lost ; a wider range of notes again causes the reflex (internal inhibition again counteracted by external inhibition).

If notes closer and closer together are used the animal eventually has difficulty in distinguishing them and when faced with the task of doing so may display symptoms very like those of human anxiety neurosis. In his later years Pavlov devoted much work to these experimental neuroses. It was the observation by Fulton and his associates in America that a chimpanzee with such a neurosis lost her anxiety after removal of the frontal lobes that led to the development of the operation of prefrontal leucotomy by the Portuguese surgeon, Moniz.

Discriminatory conditioned reflexes have been very widely used by experimental psychologists for testing the performance of sense organs in animals. If a dog can be trained to salivate to one note and not to another half a semitone below it, it obviously has pitch discrimination at least as good as that. It is important to note that the reverse does not follow. In this type of experiment the animal has not only to distinguish the two stimuli but also to remember the first stimulus for several minutes. Dogs happen to have an excellent memory for pitch, but many experiments claiming to show that various animals do not have colour vision may merely mean that they have no colour memory.

Pavlov's work, which until recently gave the direction and, indeed, the language to so much of the thought and the research in physiology and medicine in Russia and the countries in her sphere of influence, was important because it was the first experimental investigation of higher nervous function in animals. He was careful not to use subjective words like perceive, desire, etc., in describing the results of the experiments. A great deal was learnt about the rules of habit formation in the cerebral cortex, but, of course, it was not and has not yet been possible to get at the neuronal mechanisms involved.

The Electrical Activity of the Cerebral Cortex

Nerve impulses can be detected arriving in the projection areas of the cortex and leaving the motor area, but what goes on in between is mysterious. When recording electrodes are placed on the surface of the brain potential oscillations are recorded all the time, even when no

sensory impulses are arriving, unless the animal is deeply anæsthetised. In a normal conscious man such brain waves can be detected through the skull by electrodes resting on the scalp; the record is called the electroencephalogram (E.E.G.). Over most of the head it is usual to find regular more or less sinusoidal waves, some 50 microvolts (5×10^{-5} volts) in size with a frequency of about 10 cycles per second. These waves, known as the alpha rhythm, were first described by Hans Berger in 1929. The alpha rhythm is only present when the eyes are shut; with the eyes open it is replaced by lower voltage, irregular activity (Fig. 14.19). It appears that the alpha rhythm is characteristic of an awake but inattentive brain. Thus even with the eyes closed the alpha rhythm may disappear if the subject's attention is engaged, if he tries to do mental arithmetic, or if he is unexpectedly touched. Normally

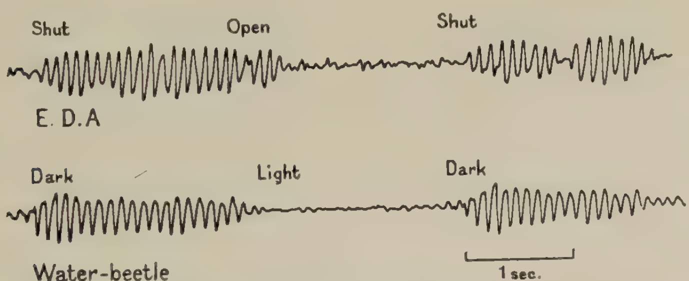


FIG. 14.19. The electroencephalogram of a human subject (E.D.A), recorded with electrodes on the scalp, showing the prominent α rhythm at about 10 cycles/sec with the eyes shut, and its disappearance when the eyes are open.

The lower record shows the very similar rhythm obtained by leading from the optic ganglion of a water beetle in the dark, which again is blocked by allowing light to enter the eye. (Adrian and Matthews, *Brain*, vol. 57, p. 355, 1934.)

it is difficult to be inattentive with the eyes open, for vision is the dominant sense, but if the scene is made uninteresting, for example by wearing strong spectacles which blur everything, or by looking at a blank screen, the alpha rhythm returns. If inattention goes so far that the subject falls asleep, the 10-per-second rhythm disappears and is replaced by irregular higher voltage waves of much lower frequency, 1–3 per second.

The alpha rhythm must be due to the synchronised beating of very large numbers of cortical neurones, for if few neurones were involved or if large numbers were beating asynchronously at different frequencies the resultant potentials would not be large enough to detect through the skull. The detailed mechanism of the alpha rhythm is not understood, but it seems to be an example of a general tendency to synchronised rhythmical oscillation that is shown by many large masses of nerve cells when they are not doing anything in particular. Remarkably similar waves were recorded by Adrian from the optic ganglion of a

water beetle, and, like the alpha rhythm, they disappeared when a light was shone into the eye (Fig. 14. 19).

Nothing very definite can be made out in the cortex when attention is aroused by the arrival of a sensory message. Leading directly from the cortex in animals has shown that the synchronised activity of large regions is broken up and replaced by an indecipherable shifting pattern of waves of varying amplitude and frequency.

The human electroencephalogram is important in medicine. During an epileptic fit very large waves occur all over the brain. Many epileptics continue to show some abnormal waves in between attacks, and their discovery is often useful in diagnosis. Distinctive abnormalities also occur around cerebral tumours and other lesions, and assist in their location.

The cerebellar cortex also shows spontaneous rhythmical waves, but of a much higher frequency, some 250 per second. They cannot be picked up by electrodes on the skin in man.

Consciousness and Sleep

Throughout this chapter it is taken for granted that conscious happenings in the mind, sensations, emotions, voluntary efforts, etc., are associated with definite physical events in the brain. This is a reasonable working assumption until it is shown to be wrong and the resultant mixture of physical and psychological terms it encourages (as when we speak of a tract of nerve fibres conveying sensations of pain) saves space and is not likely to mislead unless it is taken out of context. The nature of the interaction between mind and brain is so far a matter for philosophers, but physiologists have something to say about what parts of the brain are involved in conscious life.

A distinction must be drawn between the state of consciousness, whether we are conscious or unconscious, and the content of consciousness, what we are conscious of. The content of consciousness appears to depend mainly on the normal activities of the cerebral cortex. We are ordinarily aware of what we see and hear and feel and of what we are thinking and doing, and so on. As described in previous sections it looks as if when the region of cortex dealing with one of these activities is lost that side of consciousness drops out. It is apparently wrong to picture the conscious self as a little man sitting in a room on the top storey answering telephones when he's rung up by the various sensory areas. This puts the difficulties one stage further off, but there doesn't appear to be such a stage. The sensory areas seem to be straight through to consciousness without having to ring up, and when one of them is lost the conscious self is that much diminished. Were all the cortex removed we should presumably be conscious only of the crude skin sensations attributed to activity in the thalamus ; and with the thalamus gone too, of nothing. Thus the content of consciousness seems to depend mainly on how much of the cerebral cortex is there, or rather on how much of it is in a normally responsive state. It does not respond normally to sensory impulses arriving if an anæsthetic

has been given or if an epileptic fit is in progress, and at these times the subject is unconscious.

In ordinary life whether the cortex has the responsiveness of the normal waking state or whether it is unresponsive and we are asleep seems to be determined by what goes on in the hypothalamus and neighbouring parts of the midbrain and thalamus. It has been known for a long time that tumours growing in this region may cause pathological sleepiness and that mechanical interference around the midbrain at operation under local anæsthetic is liable to cause a rapid lapse into unconsciousness. On the other hand large pieces of cortex can be removed at operation under local anæsthetic without any apparent alteration in mental state. Probably it is for the same reason that a punch on the jaw which transmits the shock to the base of the brain is a much more effective way of knocking a man out than hitting him on the top of the head.

The precise location and mode of action of these lower centres is at the moment the subject of much research. Magoun and his followers urge the claims of the reticular formation of the brain stem to be the region responsible for keeping the cortex awake. The principal evidence is firstly that lesions destroying the grey matter around the aqueduct of the midbrain in animals leads to somnolence. This is not due to interference with the sensory tracts running to the thalamus, for it occurs when these are not damaged ; conversely lesions interrupting the sensory tracts but not the reticular formation do not cause sleepiness. Secondly, electrical stimulation of this region in anæsthetised animals converts the electrical activity of the cerebral cortex from the pattern characteristic of sleep to that of wakefulness (see p. 449). Similar " arousal reactions " cannot be obtained from elsewhere in the midbrain.

There is evidence, then, for the view that normal wakefulness depends on the action of neurones in or near the reticular formation. They, as it were, turn the cortex on. How they and the cortex are turned off again is not clear. In the nineteen-twenties Hess showed elegantly that stimulation of the lower part of the thalamus could cause sleep. He used freely moving conscious cats with electrodes chronically implanted into the thalamus. Appropriate stimulation caused a cat to look sleepy, circle round, lie down and go to sleep, exactly like a normal cat going to sleep. There was no question that the cat was merely knocked unconscious by the current. These experiments suggest that falling asleep may be an active process. There is evidence in man too that falling asleep is more than something that happens when all sensory inputs are reduced to a minimum, although this assists. In the mysterious disease, narcolepsy, the patient is seized with irresistible drowsiness during the day, perhaps several times a day, and rapidly falls asleep. After a minute he can be aroused and is once more normal.

CHAPTER 15

THE AUTONOMIC NERVOUS SYSTEM

The autonomic (involuntary or vegetative) nervous system is often defined as that part of the peripheral nervous system which is independent of the control of the will. The designation of the system as "autonomic" or "involuntary," stresses this essential feature. The definition, however, like nearly all attempts at definition, is not strictly correct. Reflexes like the knee jerk are involuntary, but do not involve autonomic pathways. On the other hand, by recalling emotions or sensations, the will may exercise more or less control over smooth muscle and glands and in doing so stimulate autonomic fibres. There are, in addition, persons who can control one or another of the autonomic functions. They are able, by effort of will, to slow or to quicken the heart beat, to contract the smooth muscle of the skin or to constrict the pupil.

For our purpose, the autonomic nervous system is best defined as the efferent pathway to the viscera, including all nerve cells and fibres through which impulses are sent from the central nervous system to glands, smooth muscles and heart, that is, all the efferent fibres in the body except those to the striped (voluntary) muscles. The definition stresses two points, (*a*) the efferent character of the system, and (*b*) the fact that we are dealing with a peripheral nervous system. Afferent (sensory) fibres are usually excluded from the autonomic nervous system. When we use the term autonomic nervous system, however, in this restricted sense, we have to realise that most of our so called autonomic nerves, such as the splanchnic, the vagus, or the chorda tympani, are mixed nerves containing autonomic (efferent) as well as afferent (sensory) fibres.

Autonomic nerve fibres emerge from the central nervous system, mostly in the motor roots of the spinal cord and in some of the cranial nerves. There are two major divisions of the autonomic nervous system, the *sympathetic nervous system* which emerges from the thoraco-lumbar spinal region, and the *parasympathetic nervous system* which emerges from the cranial and sacral regions. Many organs in the body are influenced by both sympathetic and parasympathetic systems, usually in opposing directions, but the definition of the two systems derives from the anatomical connections with different parts of the central nervous system, not with the way in which physiological function is changed.

There are some organs to which the double antagonistic innervation does not apply. For example, most of the systemic blood vessels are supplied by sympathetic vasoconstrictor fibres only, and centrally induced vasodilatation is brought about by inhibition of vasoconstrictor tone. Some vessels in skeletal muscles are supplied with vasodilator

fibres, but these, also, belong anatomically to the sympathetic system (p. 467 below). There is no parasympathetic innervation of the smooth muscles of the upper eyelid, of the hair muscles in the skin, of the uterus, of the sweat glands or of the adrenal medulla. Even when a tissue is provided with a dual antagonistic innervation, a too simplified conception of this antagonism is misleading. When light falls into the eye the pupil constricts as a result of parasympathetic impulses. When the eye is in the shade the pupil dilates, not as a result of stimulation of the sympathetic fibres to the dilator muscle of the pupil, but because of diminished discharge in the parasympathetic system. On the other hand, the dilated pupil of a frightened person is mainly a sympathetically stimulated pupil and dilatation can occur despite the fact that bright light may shine into the eye.

To understand the difference in the working of the different parts of the autonomic nervous system the anatomical arrangements have to be considered. Efferent nerves in general can be subdivided into somatic and autonomic. The somatic nerves pass as medullated nerve fibres from the ventral horn cells in an uninterrupted pathway to the skeletal muscles. In contrast, the autonomic peripheral pathway consists, with one exception, of two neurones. This is illustrated in Fig. 15. 4 (p. 468). Impulses emerging from the central nervous system into autonomic fibres thus have to cross a *ganglionic synapse* on their way to the periphery. The first neurone originates in a cell in the central nervous system and the axon (the *preganglionic fibre*) terminates at a synapse near a cell in a ganglion. The ganglion cell with its axon (the *post-ganglionic fibre*), constitutes the second neurone which terminates at the peripheral effector structure. A preganglionic fibre may traverse one or several ganglia without entering into synaptic junction with a ganglion cell ; each fibre may form a great number of synapses in a given ganglion or give off collateral branches terminating around the cells of different ganglia. By using *nicotine*, Langley was able to discover the endings of preganglionic fibres and the origin of the postganglionic ones. Nicotine has no action on nerve fibres, but stimulates, and later paralyses the ganglion cells. If it is painted on an autonomic ganglion, therefore, the nerve cells in the ganglion will be excited and give rise to impulses passing along the postganglionic fibres, thus indicating the origin of a postganglionic fibre to a given peripheral structure. Later on, these cells become paralysed and can no longer be excited when their preganglionic fibres are stimulated. This block will not occur when fibres merely traverse a nicotinised ganglion without entering into synaptic junctions. It is, therefore, possible by stimulating fibres proximal to a ganglion, painted with nicotine, to find out if they only traverse the ganglion or form synapses with its nerve cells.

The position of the cell stations or synaptic junctions in the autonomic path varies. Preganglionic fibres may be relatively short, relaying in ganglia near the vertebral column (vertebral or lateral ganglia), to long postganglionic fibres. Or the postganglionic fibres may be short and the long preganglionic fibres may terminate around ganglion cells

situated within the tissue of the innervated organ (terminal or peripheral ganglia). Or the cell stations may have intermediate positions (collateral or prevertebral ganglia). There is this general difference. Parasympathetic preganglionic fibres relay in peripheral ganglia or in collateral ganglia (ciliary, sphenopalatine, submaxillary or otic ganglia) situated near the innervated tissue. Terminal ganglia, however, are the exception in the sympathetic path. Its main cell stations are the vertebral ganglia forming the paired sympathetic chains with their adjoining cervical ganglia in the neck. In addition there are many cell stations in collateral ganglia, like the cœliac and the mesenteric ganglia. There is one exception. The autonomic fibres to the adrenal medulla do not pass a cell station on their way to the periphery. The medullary cells and the sympathetic ganglion cells have a common origin. Both are probably derived from the same primitive masses of neuroblasts, but have followed different paths in their differentiation. The innervation to the adrenal medulla consists of preganglionic fibres of the sympathetic system, the postganglionic fibres, so to speak, having been converted into a gland of internal secretion.

Since the preganglionic sympathetic fibres are medullated when they emerge from the lateral horn cells of the cord into the ventral roots and pass as fine filaments to the sympathetic chain, these filaments have a whitish appearance (white rami communicantes). The fibres usually lose their myelin sheath near or at the ganglia ; thus the connecting filaments containing the non-myelinated postganglionic fibres which are sent back to the spinal nerves of the trunk and limbs have a more greyish appearance (grey rami communicantes).

The term *intermediate ganglia* is given to sympathetic nerve cells which are not located in the ganglia of the paired sympathetic chains, but in the rami communicantes, often in close proximity to the motor nerves. They are of practical importance in surgery of the sympathetic nervous system. For instance, an operation of thoraco-lumbar sympathectomy, *i.e.* removal of the lower ganglia of the sympathetic chains, does not lead to complete sympathetic denervation of the skin in the lower part of the body.

The ganglia are distributing centres, in which impulses from a single preganglionic fibre may be relayed to many (twenty or more) postganglionic fibres. In the parasympathetic division the relays are situated in or quite near the innervated tissue and this limits the spread of the impulses beyond a restricted area. The position is different in the sympathetic system. Here the anatomical arrangement clearly favours diffuse distribution of the nerve impulse over wide areas. In addition, on stimulation of the sympathetic, hormones (adrenaline and noradrenaline) are secreted from the adrenal medulla into the blood stream, mimicking many of the sympathetic nerve effects, thus emphasising the fact that the organism is not concerned with a limitation of sympathetic effects to restricted areas. The anatomical arrangements of the two divisions reflect the fundamental difference in the function of the two systems.

It is possible that some of the parasympathetic ganglia are weak automatic centres, independent of the central nervous system, from which impulses originate continuously. Such function is attributed to the cells of the nerve plexus in the wall of the digestive tract. To a lesser degree it may be a more general property of parasympathetic ganglia.

The adrenal medullary cells secrete adrenaline and noradrenaline as a result of the release of acetylcholine from the pre-ganglionic nerve endings (see p. 469). We must emphasise again the distinction between transmitter substances and hormones. The former are released from nerve endings and then modify or initiate the activity of cells in close proximity : the latter are secreted from gland cells, released into the blood, and act on cells which may be quite remote and in entirely different organs.

The Function of the Sympathetic System

The Emergency Function of the Sympathico-adrenal System. Removal of the paired sympathetic chains with their outlying ganglia, as far as it is technically possible, is compatible with life. Sympathectomised animals show, in fact, no signs of deficiency if kept in sheltered conditions, but when exposed to extreme cold, oxygen lack, carbon dioxide increase, hypoglycæmia, hæmorrhage or anæsthesia they may succumb earlier than control animals. The sympathetic innervation fulfils an important function in making the animal fit for states of emergency. There is a widespread discharge of impulses in the sympathetic system in states of physiological stress, during severe muscular work, in situations of danger, in extreme temperatures, asphyxia, hæmorrhage, under strong emotions such as fear or rage, or when in pain. The discharge affects also the fibres to the adrenal medulla leading to an output of adrenaline and noradrenaline, as is illustrated in Fig. 11. 3, p. 322. This widespread discharge has been likened to a reflex action of the organism with the purpose of strengthening its powers of defence and producing those changes necessary for preparing the organism for " fight and flight." It is in this connection, that Cannon referred to the emergency function of the sympathico-adrenal system. Sympathectomised animals show signs of deficiency in many of these adverse circumstances, but the degree of deficiency varies in different species. The capacity for strenuous muscular exercise is definitely decreased in cats but not in dogs which remain excellent fighters.

The effects of sympathetic stimulation are easily understood and remembered, when seen in the light of a protective mechanism for emergencies. The dilatation of the pupil (contraction of the radial muscle of the iris) protrusion of the eye or exophthalmos (contraction of the smooth muscle at the back of the eye) and opening of the palpebral fissure (contraction of the smooth muscle fibres of the levator palpebræ) increase the perception of light. In animals an alarming appearance is produced by bristling of the hairs of the back and tail. Of this effect " goose flesh " alone has survived in man. Broncho-dilatation decreases the resistance to the passage of air into the alveoli of the lung. The movements of the digestive tract are inhibited and the sphincters contract. Glucose is mobilised from the liver. The spleen contracts

and ejects its store of red blood cells. Fatigue in skeletal muscle may be counteracted. The heart beats more strongly and more frequently (Figs. 2. 5 and 2. 12, pp. 43 and 55) ; the coronary arteries are dilated. Vaso-constriction occurs in the systemic vessels, mainly in the splanchnic area (Fig. 2. 5) and in the skin which becomes pale. In the skeletal muscles, however, the blood vessels dilate, thus shifting the flow of blood from regions where it is not urgently needed to the active tissues. The redistribution of blood may occur with little or no rise in arterial blood pressure (Fig. 2. 12). No useful purpose would be served if the circulatory effects of sympathetic stimulation, as well as of adrenaline, consisted in an increased activity of the heart in order solely to eject the blood against a greater peripheral resistance. The main effect is redistribution of the blood volume with increased circulation rate. On the other hand, the vasoconstriction following severe hæmorrhage tends to keep up or restore an effective arterial blood pressure by adapting the vascular bed to the reduced blood volume.

Adrenaline and noradrenaline are apparently not always secreted in a constant proportion. They may be secreted by different cells which can be selectively stimulated from the central nervous system. Insulin hypoglycæmia, for example, increases the proportion of adrenaline secreted, while most other ways of evoking central stimulation of the adrenal medulla produce a preponderance of noradrenaline in the secretion. It is interesting to note in this connection that adrenaline has a much more powerful hyperglycæmic effect than has noradrenaline.

Cannon points out that a general sympathico-adrenal discharge may be harmful unless transformed into action. Heart and circulation may be worked just as hard from an armchair as from a rower's seat. " If no action succeeds the excitement and the emotional stress—even worry and anxiety—persists, then the bodily changes due to the stress are not a preparatory safeguard but may be in themselves profoundly upsetting to the organism as a whole."

The Efferent Pathway of Reflexes to Organs Widely Distributed in the Body. The sympathetic system may act as a unit in conditions of physiological stress, but this is one aspect only of its function. It may also act as the efferent pathway for reflexes in which the blood vessels, sweat glands and hair muscles are the effector organs, widely distributed in the body. Sympathetic fibres are the sole connections between the vasomotor centre in the brain and the blood vessels ; regulation of the calibre of the vessels is brought about by increased or decreased sympathetic discharge, the parasympathetic taking no part. It is true that the latter system contains vasodilator fibres to some tissues such as the salivary glands ; these fibres, however, are not activated for the purpose of circulatory readjustments but for a specific organ function, salivation. Without a sufficient supply of fluid to the glands, salivation would not continue ; accordingly the pattern of the salivary reflex incorporates a localised vasodilatation mediated by parasympathetic nerves. The regulation of heat loss

through the temperature regulating centres (Chapter 20, p. 589) is almost entirely dependent on the sympathetic system, whether by control of the calibre of the blood vessels in the skin and thus the skin temperature, control of the evaporation of sweat or, in animals, control of the erection of hair or feathers.

Sympathetic discharge to certain tissues is continuous. The heart, the arteries, arterioles, capillaries and probably venules, and the smooth muscles in and around the eye, are kept in a state of continuous although varying tonic contraction as a result of their sympathetic innervation. When this is interrupted, the blood vessels dilate, as was first shown by Claude Bernard when he cut the cervical sympathetic nerves on one side of a rabbit ; the vessels of the external ear on the denervated side dilated and the skin temperature rose. When the sympathetic innervation to the eye is interrupted, the pupil contracts, the eye sinks into its socket (enophthalmos) and the upper lid droops (ptosis), giving the eye a sleepy appearance. The smooth muscles of the hairs, the sweat glands, the digestive tract and the medulla of the adrenal gland, on the other hand, receive sympathetic excitation only in special conditions, *e.g.* those of an " emergency."

It is not surprising, therefore, that different parts of the sympathetic system may act separately from each other and even antagonistically. The following instances will illustrate these points. (1) Emotional blushing is the result of inhibition of sympathetic constrictor tone of the skin vessels. This inhibition is limited and usually does not spread over the vessels of the whole body. In women who blush frequently and vividly, the " blush area " is usually confined to the face and to the V-shaped area in the neck, areas of skin exposed to sunlight by the cut of modern dress. (2) Sweating, limited to the skin around the lips and nose, may be evoked by gustatory stimuli such as chewing spicy foods. (3) When there is a sufficient rise in the environmental temperature, large areas of the skin become flushed and sweat beads appear. Again the flushing is due to reflex *inhibition* of sympathetic vasoconstrictor tone in the skin but the secretion of sweat is the result of *excitation* of the sympathetic secretory fibres to the sweat glands. When the vasoconstrictor fibres are excited at the same time as the secretory fibres, as in extreme fright, " cold sweat " appears, the sensation of cold being brought about by the restriction of the blood flowing through the skin.

One of the main functions of the sympathetic system is its *rôle* in preserving constant internal conditions, the preservation of what Claude Bernard called the " milieu intérieur." The sympathetic system is in part responsible for man's great adaptability to life in different surroundings and for the conservation of his " inner climate " which he carries about with him. The constant changes in the distribution of the circulating blood volume to adapt the organism to changed environmental conditions and to the changing demands created by muscular activity are brought about, as far as nervous mechanisms are involved, through the sympathetic system.

The Function of the Parasympathetic Nervous System

Unlike the sympathetic system with its widespread discharge the parasympathetic system is the main efferent pathway for those reflexes which are more localised and usually influence single organs without affecting others. These reflexes are abolished when the parasympathetic pathway is interrupted ; for example, in the eye the pupillary reflexes to light and near vision (Chapter 18, p. 533) are no longer obtained. When the parasympathetic pathway to the salivary glands is interrupted neither the presence of food in the mouth nor its sight or smell will induce salivary secretion. The reflex secretion of gastric and pancreatic juice and of succus entericus are dependent on the integrity of the parasympathetic fibres in the vagus nerve, stimulation of which in addition, causes increased bile flow and increased activity of the walls of the digestive tract and inhibition of its sphincters. Cutting the parasympathetic fibres to the lachrimal glands abolishes reflex lachrimation. A continuous discharge is exerted through the parasympathetic fibres in the vagus upon the heart's action, as shown by the fact that the heart-rate in man may double when the vagal inhibition is removed, as after atropine. Vagal tone is weak at birth : in a new-born baby, atropine will increase the pulse rate only from about 140 to 160 per minute. The vagal tone to the heart in man is influenced continuously by many reflexes. The significance of this tone depends on the effect of heart rate on mechanical efficiency (Chapter 1, p. 23) ; the heart uses less oxygen to perform a given amount of work when it is beating slowly than when it is beating quickly.

The parasympathetic fibres from the sacral division are the efferent pathway for the reflex contraction of the urinary bladder and inhibition of its internal sphincter in the micturition reflex. There is no sympathetic control of bladder activity, although sympathetic nerves regulate the blood flow in the bladder muscle. The contraction, produced by stimulating the sympathetic nerve of the muscle of the ureteral orifices and of the trigonum, is linked not with the micturition reflex but with the sex function. Section of the hypogastric nerves which contain the sympathetic fibres does not interfere with micturition whereas the bladder becomes paralysed after section of the pelvic nerves.

The cranial division of the parasympathetic contains vasodilator fibres to the salivary glands and tongue and the sacral division contains similar fibres to the erectile tissue of the external genitalia. The main rôle of these vasodilators is, as mentioned before, linked with the specific functions of these organs, salivary secretion and erection of the generative organs respectively and not with general circulatory readjustments. Ejaculation is dependent on the integrity of the sympathetic system ; its removal causes impotence in the male. Thus both divisions of the autonomic nervous system are involved in the mechanism of coitus (Chapter 10).

TABLE 15. 1

Summary of the Effects of Stimulation of the Sympathetic and Parasympathetic Nerves

Organ	Sympathetic	Parasympathetic
Glands		
Sweat	Secretion	No innervation
Salivary, Gastric, Intestinal and Pancreas (acini and islets)		Secretion
Liver	Glycogenolysis	Increased bile flow
Lachrimal		Secretion
Smooth muscles		
of bronchi	Relaxation	Contraction
of œsophagus	Relaxation ; usually contraction of cardiac sphincter	Contraction ; relaxation of cardiac sphincter
of stomach	Usually relaxation	Contraction
of intestine	Relaxation	Increased tone and motility.
of eye		
iris	Midriasis ; contraction of dilator pupillæ	Miosis ; contraction of constrictor pupillæ
ciliary	No innervation	Contraction
internal anal sphincter	Contraction	Relaxation
detrusor of urinary bladder	Relaxation	Contraction
trigone and sphincter of urinary bladder	Contraction	Relaxation
vasa deferentia, seminal vesicles and prostate	Contraction (ejaculation)	No innervation
uterus	Relaxation ; contraction when pregnant	No innervation
Blood vessels of		
salivary and lachrimal glands	Constriction	Dilatation
abdominal and pelvic viscera	Constriction	No innervation
external genitalia	Constriction	Dilatation (erection)
skin	Constriction	No innervation
skeletal muscles	Constriction. Dilatation during activity	No innervation
coronary system	Dilatation	Constriction (?)
Heart		
frequency of beat	Increased	Reduced
conduction of impulse	Quickened	Slowed
auricular contraction	Strengthened	Weakened
ventricular contraction	Strengthened	No innervation in mammals

Relations between the Autonomic Nervous System and the Central Nervous System

A man's conscious activities largely consist in controlling his skeletal (" voluntary ") muscles in response to information received through his " special senses " (chiefly vision and hearing) and controlled by his central nervous system. But he cannot do this properly unless

his " auxiliary machinery "—cardio-vascular system, gastro-intestinal tract, etc.—is also operating properly, controlled in an appropriate manner by his autonomic nervous system ; this, accordingly, is not really autonomous, as the name implies, but co-operates with all the other parts of the nervous system.

Some examples of this have been given in previous chapters. If the spinal cord is severed from " higher " parts of the nervous system, defæcation and micturition are controlled according to the distension of the rectum and bladder respectively. Gunshot wounds may sever completely the spinal cord in men. If this occurs in the cervical or upper thoracic region, disconnecting the sympathetic system from the higher centres, distension of the bladder causes a large reflex rise in blood pressure. If the section through the brain stem is such as to allow the medulla to remain connected with the spinal cord—as by decerebration, or by a gunshot wound in the lower thoracic region of the spinal cord, the blood pressure is well regulated and there is no rise when the bladder is distended. The baroreceptors are connected to the vasoconstrictor fibres of the sympathetic system through the vasomotor centres (Chapter 2, p. 42), and are able to control the cardio-vascular system. In the whole normal animal, with intact nervous system, these and other " centres " are subjected to overriding influences from the rest of the nervous system, and their efferent discharges blocked or enhanced.

Little is known about this finer central integration which is necessary for keeping the internal environment constant. The most important structures through which it is exerted, however, lie in the **hypothalamus.** As mentioned in Chapter 11, the internal secretions of the adenohypophysis (anterior pituitary body), and hence those of many other parts of the endocrine system concerned in stabilising the internal environment, are controlled through the hypothalamus. Here also are the " osmoreceptors " which control drinking (Chapter 7) and water diuresis through the neurohypophysis (posterior pituitary body) (Chapters 9 and 11), the centres controlling hunger and appetite (Chapter 7) and those controlling body temperature and the loss of heat (Chapter 20). Electrical stimulation of the hypothalamus, through implanted electrodes in unanæsthetised animals, results characteristically in excitation of the sympathetic system with rise of blood pressure, dilatation of the pupils, erection of hairs and inhibition of gastro-intestinal movements and secretion ; but parasympathetic effects may also be obtained—contraction of the bladder and increase in gastro-intestinal movements—according to the exact position of the stimulus and form of stimulating current. Stimulation of an appropriate area, may put the animal into a rage, snarling and biting, with staring eyes, hairs on end and general excitation of the sympathetic system, with perhaps urination and defæcation. A similar condition of " sham rage," ill-directed and short-lived, is produced by quite harmless stimuli in animals whose forebrains have been destroyed, leaving the hypothalamus intact.

Some structures in the hypothalamus, therefore, seem to be concerned in producing an abnormally irritable and aggressive type of behaviour, together with excitation and inhibition of many parts of the autonomic nervous system. These structures are normally held in check by higher centres which lie in the rhinencephalon, or " olfactory brain ": this consists of structures which form a kind of arch, or " limbus " round the rostral brain stem and interhemispheric commissures, and is better called the " limbic system " since there is no good evidence that it is concerned with the sense of smell. The most important of these structures, in relation to emotion and temperament, appear to be the *amygdala.* If, instead of the whole forebrain, only the neocortex is removed, leaving the amygdala intact, cats become abnormally placid and show no signs of anger even when ill-treated ; if the amygdala are then destroyed bilaterally, the cats become savage and malevolent. The evidence, however, is conflicting, since in other series of experiments, on cats and monkeys, removal of the amygdala has made the animals unusually placid ; wild Norway rats also, ordinarily untameable, become gentle. Rather similar conflicting results have followed attempts to improve the condition of assaultive psychotic patients by making lesions in the amygdala The organisation of the limbic system is obviously very complicated, and much remains to be discovered.

Chemical Transmission

The chemical transmission of excitation from motor nerves to skeletal muscles has been discussed in Chapter 13 and its intimate mechanism is most fully understood at these junctions. But historically, the theory of chemical transmission of the nervous impulse arose at the beginning of this century in connection with the autonomic system, to explain the striking similarity between the actions of adrenaline and of sympathetic nerve stimulation, on the one hand, and between the actions of drugs like pilocarpine and muscarine and of parasympathetic nerve stimulation, on the other hand.

Direct experimental evidence in favour of the theory was first produced by Otto Lœwi in 1921. Fig. 15. 1 shows a modification of his original method. Two frog's hearts are supplied with Ringer's solution from the same reservoir, the liquid being mixed by the pumping action of the hearts. On stimulating the vagi to the first heart, it is inhibited and may stop beating. When stimulation ceases and the heart starts beating again, a slight but definite inhibition occurs in the second heart which is connected with the first heart only by the Ringer solution. A substance must have been released into the liquid during stimulation of the vagus which on reaching the other heart causes the vagus-like effect. On stimulation of the sympathetic accelerans fibres to the heart an " accelerans substance " is correspondingly released.

Similar experiments with stimulation of the vagus to the mammalian heart perfused with blood failed for a long time to demonstrate the existence of a " vagus substance," for the following reason. We now know that the substance released from the parasympathetic vagus

nerve endings is **acetylcholine.** Once released, this is quickly hydrolysed by an enzyme *cholinesterase* into choline and acetic acid, both pharmacologically inert substances in comparison with acetylcholine. Choline has, in fact, actions like acetylcholine, but only if given in concentrations several thousand times as great. The amounts of choline set free, therefore, are too small to produce reactions and the hydrolysis may be regarded as an effective mechanism of inactivation. In warm-blooded animals the enzyme will have acted usually before the acetylcholine has had time to enter the capillaries. By the use of a tissue from a cold-blooded animal, and of Ringer solution instead of blood, Lœwi had avoided this danger.

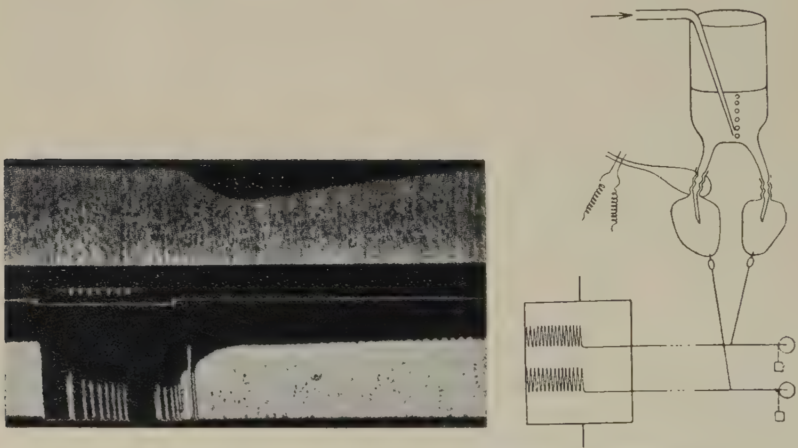


Fig. 15. 1. Demonstration of **Chemical transmission of vagus effect** on frog's heart. Modification of Lœwi's original experiment. (After Kahn.)

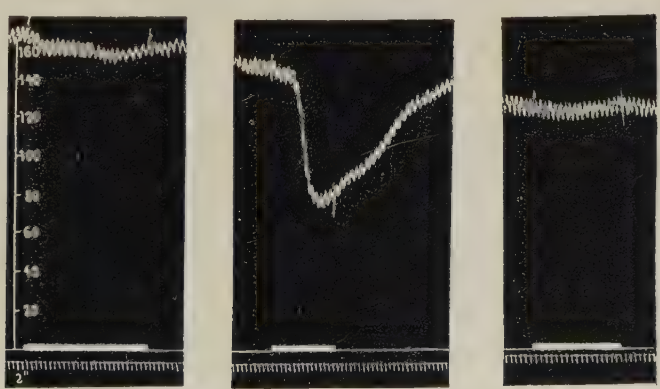
There are two kinds of cholinesterase : a non-specific pseudo-cholinesterase present in many tissues and in plasma and a true cholinesterase, sometimes referred to as acetylcholinesterase, present in nerves (and also in red blood cells), which is responsible for the destruction of acetylcholine.

The effects of sympathetic stimulation have been described as rather, but not quite, like those of adrenaline and the name **sympathin** was given by Cannon to the " accelerans ", adrenaline-like, substance or substances liberated at sympathetic nerve endings. When noradrenaline (adrenaline without its methyl group—see formulæ on p. 319) was identified in the adrenal glands, it was found to have actions resembling those of sympathin, particularly in those respects in which the actions of sympathin and adrenaline differed. Sympathin is now generally held to be a mixture of *adrenaline* and *noradrenaline* in different proportions at different sites of action. A third substance, *isoprenaline* (isopropylnoradrenaline) has been found to show those actions of sympathin in which it differs from those of noradrenaline. Isoprenaline has also been found in some mammals.

Sympathin is not as rapidly destroyed by blood as is acetylcholine,

and as a hormone, the sympathin secreted by the adrenal medulla is transported by the blood stream to the tissues on which it acts. Sympathin released from nerve endings is probably partly destroyed before diffusing into the capillaries. By the action of an enzyme in the tissues, the hydroxyl group in the ortho position on the benzene ring (see formulæ on p. 319) is methylated ; the compounds formed are pharmacologically inactive. Another enzyme (monoamine oxidase) may oxidise and deaminate the side chains. But some of the transmitter certainly reaches the blood stream in an active form and can and has been demonstrated there by its reactions on distant denervated tissues (denervation sensitises the tissues to sympathin or adrenaline). For instance, when in cats the heart, pupil and nictitating membrane (a third eye-lid present in some species) are denervated and the adrenals removed, to exclude this source of sympathin, stimulation of sympathetic fibres to other tissues will cause quickening of the heart, dilatation of the pupil and withdrawal of the nictitating membrane, all of which are typical effects of adrenaline and noradrenaline.

It would be useless to employ similar methods for the detection of the released acetylcholine. Its enzymatic destruction provides an extremely efficient safeguard against any spread of the effects of the nerve impulse and thus makes acetylcholine particularly suitable as a transmitter for the peripheral effects of the parasympathetic division of the autonomic system with its restricted localised functions. In



1 2 3

Fig. 15. 2. Evidence for the Release of Acetylcholine.

Effect on the arterial blood pressure of an anæsthetised cat of three stimulations of the chorda-lingual nerve, which contains secretory and vasodilator fibres to the salivary glands and tongue.

In 1, there is no effect, since the local vasodilatation is insufficient to affect the general blood pressure. Between 1 and 2, eserine is injected intravenously. In 2, ten minutes later, there is general vasodilatation, and a fall of blood pressure, as the released acetylcholine, escaping destruction, diffuses into the blood capillaries. Note the latency of some ten seconds, due mainly to the time taken for the blood to travel round the circulation.

Atropine is then given, and this abolishes the vasodilator action of acetylcholine. In 3, therefore, there is no depressor effect.

addition, the quick destruction will also ensure a short duration of the effect not outlasting the nerve impulse for any length of time. But how can the acetylcholine be detected if it is so quickly destroyed? This has become possible by the use of eserine (physostigmine), which prevents the hydrolysis by cholinesterase. The released acetylcholine may then escape into the blood stream and exert effects on distant organs in the same way as sympathin does in the normal course of events. This is illustrated by the experiment shown in Fig. 15. 2.

Before acetylcholine can be acted upon by the cholinesterase, it has to become attached to, or combined with, the enzyme; it is then at once hydrolysed and the enzyme becomes free again. Eserine (and related substances such as prostigmine) also combine with the enzyme and having done so, prevent acetylcholine from combining. Eserine, however, is hydrolysed slowly, if at all, and thus remains fixed to the enzyme. If, therefore, sufficient eserine molecules are available, all the enzyme molecules will, after a time, become blocked and unavailable for acetylcholine. An action such as that of eserine is called "competitive inhibition." There are certain organic phosphates which also are very potent inhibitors of cholinesterase. Some of them were originally prepared as war gases such as di-isopropylfluorophosphonate (DFP), others as insecticides such as tetra-ethylpyrophosphate (TEPP).

The amounts of acetylcholine released on nerve stimulation and available for analysis are far too small to be detected or identified by our present chemical methods. But acetylcholine has been identified chemically in extracts of the horse spleen, the human placenta and the ox brain and can be regarded as a substance occurring naturally and being formed in the body. Its identification when released on nerve stimulation is based on pharmacological methods using tissues which respond to minute doses of acetylcholine with characteristic reactions. Some tests in use for this purpose are shown in Fig. 15. 3.

(*a*) Contraction of the muscle of the body wall of the leech, the effect being greatly increased in the presence of eserine. The reaction is very sensitive; it is induced by a concentration of acetylcholine of only one part in a thousand million.

(*b*) Contraction of the rectus abdominis muscle of the frog and the sensitising effect of eserine on the action. Sensitive to about one in fifty million acetylcholine.

(*c*) Inhibition of the beat of the frog's heart. This was the first test used. The action is abolished by atropine.

(*d*) Depression of cat's blood pressure. The action is sensitised by eserine and abolished by atropine.

In each of these tests, the response to the unknown solution is matched with that of an appropriate dose of a standard solution of pure acetylcholine, and hence the apparent acetylcholine content of the unknown solution is determined. If the apparent acetylcholine content is found to be the same in all four tests, and, in addition, the unknown substance is unstable in alkaline solution and destroyed by blood in the absence of eserine, but not in its presence, the identity with acetylcholine is regarded as proved. Other choline esters produce qualitatively similar, but quantitatively different, effects on these tissues, acting relatively more on one than the other compared with acetylcholine.

In order to prove that a nerve impulse acts by the release of acetylcholine, the following three facts must be established.

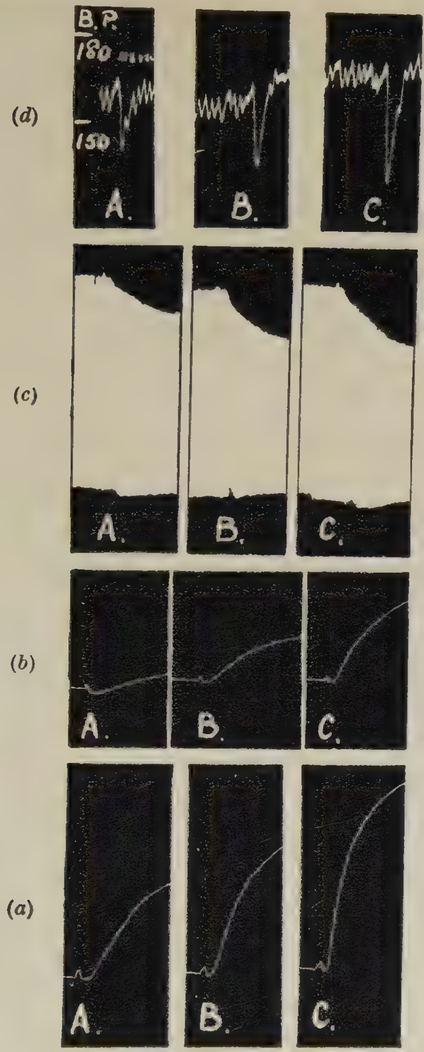


FIG. 15.3. Tests of Substance in Perfusion Fluid emerging from Veins of Stomach during Stimulation of Vagus.

From below upwards. (*a*) Eserinised leech muscle ; (*b*) Frog's rectus abdominis ; (*c*) Frog's heart[1] ; (*d*) Cat's blood pressure. In each case, A shows the effect of a suitable dose of acetylcholine, B shows the effect of a dose of the perfusion fluid, adjusted to be proportional to the dose A of acetylcholine. C is the effect of acetylcholine given in twice the concentration of A. In each of the four reactions, the effects of B are intermediate between those in A and those in C.

Concentrations of acetylcholine (A) : (*a*) $1 : 280 \times 10^6$; (*b*) $1 : 56 \times 10^6$; (*c*) $1 : 56 \times 10^6$; (*d*) 1 ml. of $1 : 40 \times 10^6$. (After Dale and Feldberg.)

[1] Owing to the slowness of the drum, the individual vertical lines representing heart beats have overlapped. The vertical distances between the upper and lower borders of the white patch nevertheless indicate the relative amplitudes of the heart beat.

(1) Acetylcholine is released into the tissues when the nerve is stimulated. The eserinised venous blood from a given organ may be collected and tested for acetylcholine, while the nerves to the organ are stimulated ; or the organ may be perfused with eserinised Ringer's solution, the nerve stimulated and the venous effluent collected and assayed.

(2) The action of injected acetylcholine must be identical with, or approximate closely to, that of nerve stimulation, although we have to take into account the fact that the method of injection does not always imitate closely the release of acetylcholine by nerve impulses.

(3) Eserine, by delaying the destruction of the released acetylcholine, must potentiate and prolong the effects of nervous stimulation. In some instances, prolonged action of acetylcholine may paralyse a reactive structure ; in that case the response to nerve stimulation should be affected similarly after eserine.

The identification of sympathin is based on similar lines of argument: similarity of the effects of sympathetic nerve stimulation with those of injection of adrenaline and noradrenaline, the actions of various drugs (as mentioned later) in modifying these effects, and pharmacological assay of adrenaline and noradrenaline in venous effluents during stimulation of sympathetic nerves.

The tests generally used are :—

(*a*) rise of blood pressure in the cat or rat after destruction of the spinal cord (to prevent vasomotor reflexes) and treatment with atropine and hexamethonium (see below, p. 469) ;

(*b*) inhibition of the contractions of the rat's uterus, rat's colon, or hen's colon.

The tests can detect the presence of 1/1,000 to 1/100 of a milligram of adrenaline or noradrenaline, but no single test will differentiate between them with certainty. The different preparations, however, have very different sensitivities to adrenaline and noradrenaline, so that by using several tests in parallel the quantity of each in a mixture may be estimated. Adrenaline and noradrenaline may also be separated from one another, and from interfering substances, by the use of paper chromatography.

A systematic analysis of the effects of different nerves soon showed that chemical transmission is not confined to the autonomic nervous system, as is made clear in Chapter 13. Moreover, the peripheral pathway of the autonomic nervous system (except that to the adrenal medulla) consists of two neurones ; transmission across the ganglionic synapses must be considered as well as that to the effector organs. All those nerve fibres or neurones from which the nerve impulses are transmitted to the next neurone, or the effector cells, by the action of acetylcholine are described as **cholinergic** ; all those from which transmission occurs by the action of noradrenaline or adrenaline are described as **adrenergic.**

Some Properties of Cholinergic and Adrenergic Nerves

The adrenergic and cholinergic nature of a nerve is not confined to the endings but is an inherent property of the whole neurone. An adrenergic or cholinergic nerve contains noradrenaline and adrenaline, or

acetylcholine, respectively, throughout the whole course of the nerve fibre ; when the nerve impulse passes along it, minute amounts of the chemical mediators are released. The difference between the fibre and the ending is only quantitative. At the endings, the process shows a local intensification to ensure transmission to a contiguous cell. No function can yet be postulated for the release along the course of the fibre.

Cholinergic nerves have the ability to synthesise acetylcholine from choline, with the aid of an enzyme *choline acetylase*, not only at their endings but along the whole course of the fibre. The synthesis is a complex process in which adenosine-triphosphate and coenzyme A, the coenzyme for acetylation (Chapter 6, p. 186) are involved. The acetylcholine so formed is not free but in loose combination with some cell constituent, probably protein ; this complex is pharmacologically inactive and resistant to the action of cholinesterase. At the nerve endings, the nerve impulse releases the acetylcholine from the bound complex, so that it becomes diffusible and pharmacologically active ; then it is at once destroyed by the true cholinesterase. About forty-eight hours after a cholinergic nerve is cut, the peripheral end loses its ability to synthesise acetylcholine and the acetylcholine store disappears; at this time the fibre is still able to conduct nerve impulses.

Another observation suggesting that the whole of a neurone is either cholinergic or adrenergic is based on regeneration experiments. When the known facts of regeneration experiments with cross-sutured nerves were reconsidered in the light of the chemical transmission theory, it became evident that cholinergic fibres could replace other cholinergic fibres and enter into functional connections with them and that adrenergic nerves could replace adrenergic ones, but a cholinergic fibre could not enter into functional connection with an adrenergic one or *vice versâ*.

The Distribution of Adrenergic and Cholinergic Neurones

This is shown diagrammatically in Fig. 15. 4. Most of the endings of the sympathetic fibres with their effector cells are adrenergic, which explains the striking similarity between the effects of noradrenaline and adrenaline and of sympathetic stimulation. The endings of the parasympathetic fibres with their effector cells are cholinergic, a fact which explains the equally striking similarity between the effects of parasympathetic stimulation and of drugs like acetylcholine, pilocarpine or muscarine. The secretory nerves to sweat glands, and the vasodilator nerves to blood vessels in skeletal muscles, are peculiar. The endings of the nerve fibres on the effector cells are cholinergic and thus appear to belong to the parasympathetic system ; but the nerves arise from ganglion cells which are situated centrally, not peripherally, and anatomically form part of the sympathetic system.

The synaptic transmission across the ganglia of the autonomic system, both sympathetic and parasympathetic, is cholinergic, and so also is the neuromuscular transmission at the motor end-plates of

skeletal muscles (Chapter 13). Two rather different types of cholinergic nerve ending must be distinguished, associated with the two classes into which the pharmacological actions of acetylcholine can be divided.

(*a*) *The Muscarine-like Action.* Muscarine is a substance of known composition closely related to choline ; it is found in extracts of a common toadstool (*Amanita muscaria*). Its effects are the same as those observed on stimulating the postganglionic parasympathetic

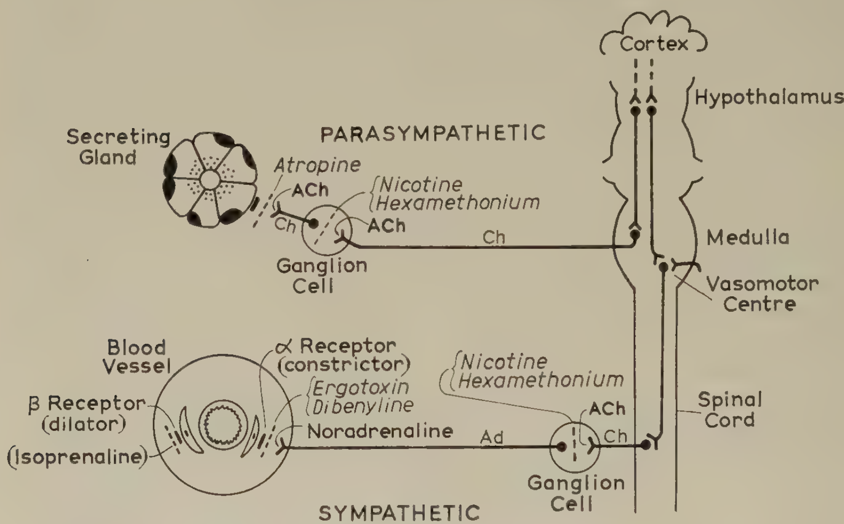


FIG. 15. 4. Diagram showing Chemical Transmission in the Autonomic Nervous System and the distribution of cholinergic (Ch) and adrenergic (Ad) neurones.

A secreting gland is taken as a typical effector organ innervated from the parasympathetic system ; and the smooth muscle in a blood vessel as a typical effector organ innervated from the sympathetic system.

At each junction the transmitter substance liberated at the nerve endings is labelled (ACh = acetylcholine) ; the receptor is shown by the thick short line ; and the site of action of blocking agents, as labelled, shown by the interrupted line. As hormones, secreted by the adrenal glands, noradrenaline acts only on the α receptors, and adrenaline, in low concentration, more on the β receptors than the α receptors ; in larger concentrations it acts chiefly on the α receptors.

The parasympathetic receptors may be excitatory, as in glands and in smooth muscle cells of the alimentary canal, or inhibitory, as in the heart. The sympathetic β receptors in the blood vessels (when present) are not innervated ; but in other kinds of receptor cell, such as the heart, they are thought to be innervated and may be excitatory.

fibres. These actions, whether induced by muscarine or acetylcholine, are abolished by atropine.

The muscarine-like action of acetylcholine is effective in the transmission to the peripheral structures from all postganglionic parasympathetic nerves, as well as from the cholinergic postganglionic sympathetic nerves ; all these nerve effects are blocked by atropine (Fig. 15. 4, upper part). After atropine, stimulation of the vagus no longer

inhibits the heart, and stimulation of the sympathetic no longer causes sweating in human beings. Nevertheless the nerve impulses still release their acetylcholine. Atropine, although abolishing the effects of nerve stimulation, has no action on the nerves or nerve endings themselves, but renders the effector structure insensitive to the action of acetylcholine, whether released or artificially applied. Many other drugs can still act after atropine. We do not know why drugs like atropine render the cells insensitive to one kind of drug and not to another.

(*b*) *The Nicotine-like Action.* Nicotine first excites and subsequently paralyses the following structures : (1) the autonomic ganglia ; (2) the cells of the adrenal medulla and (3) the motor end-plates of skeletal muscle. Similar actions can be obtained when acetylcholine is injected. The nicotine-like effects are relatively insensitive to atropine. Acetylcholine has, in addition, stimulating and paralysing effects on cells of the central nervous system.

The nicotine-like action of acetylcholine is effective in the transmission to the cells of the adrenal medulla, the nerve cells of the autonomic ganglia, and the end-plates of skeletal muscle. Transmission at these structures is not blocked by atropine, but is blocked by suitable amounts of nicotine ; the motor end-plates, also, are blocked by curarine, the active principle of the poison curare. Nicotine acts on the ganglia, and curare on the end-plates, much as atropine acts on the heart, smooth muscles or gland cells. They prevent the nerve impulses from being transmitted across the junction, but do not prevent the release of acetylcholine at the nerve endings by the incoming impulses. The action is on the ganglion cells, or the end-plates, respectively. The postganglionic fibres in the presence of nicotine, and the muscle fibres in the presence of curare, respond as usual to direct electrical stimulation.

As mentioned in Chapter 13 drugs with a curare-like action are widely used in general anæsthesia to produce relaxation of skeletal muscles. Since one cannot breathe without active skeletal muscles, this would be dangerous unless the synthetic curarine substitutes employed had a shorter duration of action than the natural product. Correspondingly, since autonomic ganglia when paralysed have as their most widespread and dramatic consequence a profound vaso-dilatation and fall in arterial blood pressure, ganglion-blocking drugs are widely used in the treatment of patients with chronic high arterial pressure (hypertension). The actions of the short-lived acetylcholine and the prolonged nicotine would be unsuitable because of their initial stimulating action. Synthetic substances such as hexamethonium bromide have an uncomplicated blocking action which is inconveniently short-lived, but many other synthetic products are now available which have a longer-lived action in blocking the autonomic ganglia in hypertensive patients.

The actions of the transmitters at the sympathetic neuro-effector junctions are illustrated diagrammatically in Fig. 15. 4 (lower part). Noradrenaline makes smooth muscle contract, for example it produces vaso-constriction and a rise of arterial pressure. This is termed the α (alpha) effect and the drug is said to act on α receptors in the muscle. (These receptors correspond with the chemically excitable structure, or the " receptive substance " at the neuro-muscular junctions, and

must not be confused with the " sensory " receptors). Isoprenaline in general produces relaxation of smooth muscle, for example vaso-dilatation and, more important therapeutically, dilatation of bronchioles. This is termed the β (beta) effect, the drug acting on β receptors. Adrenaline produces a mixed effect since it acts on both kinds of receptor, but in low (*i.e.* physiological) concentrations chiefly on the β receptors ; this action is important in blood vessels of skeletal muscles, but not in those of other organs and tissues. Adrenergic blocking agents, like ergot extracts or dibenyline, are especially effective in blocking all actions on the α receptors. Adrenaline in larger concentrations acts more on the α receptors than on the β receptors and produces vaso-constriction and rise of blood pressure ; but when administered after adrenergic blocking agents, it produces vaso-dilatation and fall of blood pressure. The effects of sympathetic stimulation on tissues other than smooth muscle, for example increased frequency of heart beat or rise in the blood sugar concentration, are regarded as β effects, partly because isoprenaline has more action on them than has noradrenaline, and partly because of their responses to adrenergic blocking agents.

CHAPTER 16

MUSCLES

THE function of muscles is to move the structures and other tissues to which they are attached or, alternatively, to prevent them from moving when acted on by other forces such as that of gravity. They do this by " contracting "; but this word is a little misleading, since the change in volume of the muscle substance is very small, any reduction in length being accompanied by thickening, as illustrated by the familiar example of the human biceps. By " contraction " of a muscle, therefore, we mean that change in length, or if the length be fixed, that increase in tension, which is the mechanical indication of its activity.

There are three kinds of muscular tissue, classified according to their detailed structures, gross and minute, their mechanical properties and the particular kind of function which they carry out. In addition, there are two other kinds of cell which can produce movement. They appear, histologically, very different from muscle cells, but it is very probable that the molecular structures by which movement is produced are essentially similar.

(1) *Striated muscles* are characterised by the transverse dark and light striations visible in the fibres under a microscope. The fibres are very long and thin, less than $\frac{1}{10}$ millimetre (10 to 100 μ) in diameter, but several centimetres in length ; each fibre may run the whole length of the muscle from tendon to insertion. They are also known as striped, skeletal, or voluntary muscle, the latter two names indicating their function in a general way. Such muscles are controlled by nerves connected directly with the central nervous system, and they effect movements such as those of a limb, most of which are under voluntary control. They are inactive in the absence of external stimulation, but contract and relax very rapidly (in a few hundredths of a second in mammals) when stimulated.

(2) *Unstriated muscles* show, in contrast with striated muscles, no microscopic transverse striations. The fibres are much smaller than those of striated muscles, being some 1 to 10 μ in diameter and about $\frac{1}{2}$ millimetre (200 to 700 μ) long ; when the whole organ or tissue contracts, the fibres must pull on each other. They are also known as smooth, plain, visceral or involuntary muscles, and are controlled by nerves of the autonomic nervous system, one set of which makes them contract and another set (usually) makes them relax (see Chapter 15). Such muscles effect contractions and relaxations of hollow organs (*e.g.* the stomach, intestine, bladder, ureter, uterus and the blood vessels), changes in which normally occur without the individual being aware of them. They usually exhibit some degree of spontaneous activity in the absence of external stimulation, and this may be expressed as a

permanent partial contraction known as "tonus," which may be increased or decreased by nervous control; or it may be expressed by periodic changes of length known as "spontaneous contractions." Thus an unstriated muscle, stretched by a small constant load, may adopt various lengths at different times, and it may be impossible to name the external influences which presumably have brought about the change; this is one of the most striking differences between visceral and skeletal muscles. Contraction and relaxation of unstriated muscle is characteristically slow, the durations being measured in seconds or minutes.

(3) *Cardiac muscle* is found only in the heart, and in many respects its properties are intermediate between those of skeletal and visceral muscles. The fibres are striated, but relatively short, less than 100 μ long and 10 to 20 μ in diameter; they are branched, and joined to several others at septa which are penetrated by fibrils, so that the whole forms a *syncytium*. Like those of many visceral muscles, but unlike those of skeletal muscles, they exhibit spontaneous rhythmic activity. The speed of movement is much greater than that of most kinds of visceral muscle, but less than that of skeletal muscles: the heart of a humming bird beats ten times a second, and the human heart can beat three times a second during exercise.

(4) *Cilia* and *flagella* are thread-like protoplasmic processes which protrude from certain specialised kinds of cell. Cilia are relatively short, their length being comparable to the size of the cell, and perform co-ordinated whip-like movements to and fro. The fluid in which they are immersed and any objects suspended in it, are thus made to move over the surface of the ciliated tissue. In mammals, the walls of the respiratory and certain other passages are lined with ciliated cells which transport mucus and particles of dust, for example. Flagella are very long compared with the size of the cell; they have wave-like movements, the waves progressing from one end to the other, and the whole cell "swims" through the fluid in which it is immersed. The tails of spermatozoa execute such movements, as already described in Chapter 10 (p. 278).

(5) "*Amœboid*" *cells* creep over surfaces by putting out processes (*pseudopodia*) in one direction and then drawing the rest of the cell into them. The white blood cells (Chapter 21, p. 610) progress by this means.

Most of what we know about the properties of muscles is derived from experiments which have been performed on them after removal from the animal. This is partly because it is then easier to handle them—for example to suspend their ends from two really rigid supports—and partly because an isolated muscle is not subject to the influence of fortuitous changes occurring elsewhere in the body. On the other hand, removal from the body involves an arrest of circulation through the muscle, and since surviving tissue requires a supply of oxygen for its maintenance in a state in which it can continue to function, isolation raises certain difficulties in technique and problems in interpretation which might be avoided in experiments on the whole animal. Chemical

reactions are retarded by a fall in temperature, and the rate of oxygen consumption of the muscles of cold-blooded animals, at room temperature, is much less than that of warm-blooded animals, at body temperature. It happens that a thin muscle, such as the frog's sartorius, can maintain its excitability for many days when isolated and placed in air or oxygen and kept moist by bathing with suitable fluids (Ringer's solution), whereas under similarly convenient conditions insufficient oxygen diffuses from the surface into the body of a mammalian muscle to keep it functionally active. A frog's heart, removed from the animal, will continue to beat for many hours if kept moist and cool, and for several days if perfused with a suitable solution. A mammalian heart, removed from the body, continues to beat, but only for a few hours even if supplied with blood. The muscles of cold-blooded animals, also, work perfectly well at 0° C., although they are then much slower than at room temperature ; frogs cannot jump when it is cold and tortoises are much less brisk in our gardens than in their native Africa. But for this reason, it is correspondingly easier to observe the time relations of the various processes by which contraction is brought about.

The properties of unstriated muscles depend on the particular organ or tissue studied, and the kind of animal from which they are derived, much more than do those of striated muscles. Being much less active in all respects, mammalian unstriated muscles survive better after isolation than do mammalian striated muscles, and much work has been done on them. But in addition, useful information has been obtained from the muscles of invertebrate animals, which behave, on the whole, in a more regular and reproducible manner. Although the conclusions drawn from experiments on the muscles of cold-blooded animals (both vertebrate and invertebrate) must not be applied *in toto* and uncritically to those of the warm-blooded animals (including man), they have not been shown to be in any way misleading.

Excitation of Muscles

Muscles can be induced to contract by the application of any of the agents which will excite nerves : pressure, heat, many chemical substances, electric currents, for example, will all excite muscles. But for experimental purposes the stimulus usually employed is electric, on account of the relative ease with which its intensity, duration, and point of application can be chosen and varied. For unstriated (visceral) muscles, it is often convenient to use one of the many drugs and hormones to which they are particularly sensitive.

The excitatory process in muscle fibres has many features in common with that in nerve fibres, as described in Chapter 12. There is an " excitable membrane " on the surface of the muscle fibre, across which there is a " resting potential " of some 30 to 100 millivolts, the interior of the muscle being electrically negative to the external solution. When this resting potential is diminished, and the excitable membrane *de-*

polarised (partially or completely) the contractile process of the muscle is set into action.

In abnormal experimental conditions, it is possible to uncouple the activation of the contractile process from the depolarisation of the membrane. If the concentration of calcium ions in the external solution is reduced sufficiently, cardiac muscle fails to contract, although the electrical changes continue, showing that the membrane is being excited ; the size of the contraction, indeed, increases with increase in the ratio of the calcium concentration to the sodium concentration, whether the calcium concentration is increased or the sodium concentration decreased (by replacement by some other suitable cation). Skeletal muscle is less sensitive to changes in the calcium concentration of the fluid outside it, and continues to contract until the calcium concentration is reduced so far that the excitable membrane becomes affected (this happens, also, in cardiac muscle if the calcium concentration is very small). It is highly probable, nevertheless, that calcium ions are necessary for contraction ; for if a calcium salt is injected intracellularly by means of a micropipette, the muscle contracts locally without excitation of the membrane. If the excitable membrane of a muscle (striated or unstriated) is permanently and completely depolarised, by immersion in an isotonic solution of potassium sulphate, for example, there is at first a maintained state of contraction, or *contracture* ; this will eventually diminish or disappear and contraction can then be initiated by increasing the calcium concentration of the external solution, or, if the muscle is unstriated, by adding acetylcholine. These substances must now be acting directly on the contractile process, since there can be little or no change in the polarisation of the membrane.

Skeletal (Striated) Muscle. If the intensity of an electrical stimulus applied directly to a striated muscle is progressively increased, the size of the contraction produced also changes progressively. The corresponding gradation of contraction when the nerve of a nerve-muscle preparation is stimulated is due to the excitation of a progressively increasing number of nerve fibres, each of which responds in an " all-or-none " manner, as discussed in Chapter 12, pp. 355. The gradation of response when the muscle is stimulated directly might also be due to a variation in the number of fibres contributing to the successive responses, different fibres having different thresholds, or it might be that each muscle fibre contracts more or less powerfully according to the strength of the stimulus.

Experiments performed by Keith Lucas in 1905 strongly suggested that the size of the contraction depends only on the number of fibres excited. Using a portion of a very small muscle (the sterno-cutaneous of the frog) containing only a few fibres, he found that on increasing the strength of the stimulus, the size of the contraction increased in steps, the number of which was equal to the number of muscle fibres. Experiments of Pratt and Eisenberger in 1919 showed that a single muscle fibre responds with an all-or-none contraction. A single fibre may be stimulated by means of a very small electrode, and the response observed by illuminating fine globules of mercury placed on the surface of the muscle ; the weakest stimulus which produces any result at all, produces a maximal contraction of the fibre. If the stimulus is then increased in strength, there is no increase in the response until enough current spreads to an adjacent fibre to excite it also. Further addition

to the strength of stimulus adds, one by one, to the number of active fibres, but no fibre can be observed to contract less fully with a weaker stimulus than it does with a stronger stimulus. Such an experiment is illustrated in Fig. 16. 1, in which the discontinuous increase of contraction—one more step for an additional fibre—can be seen. The fact that each step corresponds with one fibre depends on microscopic observation. This does not mean that a muscle fibre contracts to the same extent in all conditions ; the contraction is known to be affected by the amount of initial stretch (see p. 487, below), by fatigue, and by the composition of the fluid bathing it, for example. It means only that when the strength of the stimulus is varied continuously, a single fibre contracts as fully as it can in the circumstances, or not at all. A whole muscle which contains many fibres gives a graded contraction

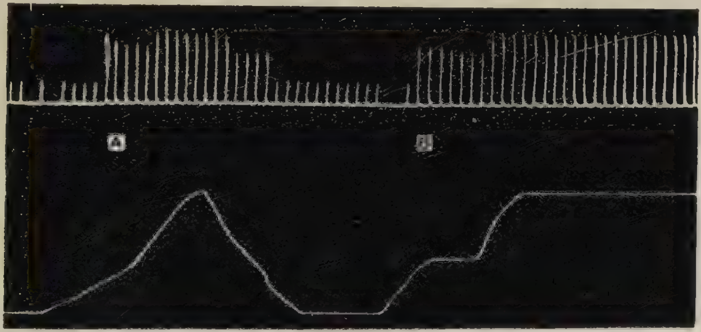


FIG. 16. 1. Response of single muscle fibres stimulated with pore electrodes. (Pratt and Eisenberger.)

The upper record shows the deflections of a mercury droplet on the surface of a frog's sartorius muscle, stimulated with successive break shocks by an electrode 7μ in diameter. Upward movement of the lower line indicates the approximation of the secondary to the primary of an induction coil. Note that graded increase of intensity of the stimulus produces either no effect on the response, or an abrupt increase corresponding with the excitation of an additional fibre.

to a graded excitation since different fibres have different thresholds, and the fraction excited depends on the strength of the stimulus.

If a pair of electrodes is placed on a striated muscle, such as the frog's sartorius in which the fibres are straight and parallel with one another, and connected to an amplifier and oscilloscope, a diphasic electric response will be observed whenever the muscle contracts. This shows that there is a propagated wave of excitation along the muscle fibres (compare Chapter 12, p. 351). If a micro-electrode is inserted into a muscle fibre, and the potential difference across the excitable membrane is recorded, it is found to reverse on excitation, being negative to the outside solution when at rest, and becoming momentarily positive when the fibre is excited. The propagated disturbance, in fact, has all the properties of the propagated disturbance in a nerve fibre, and the magnitude of the action potential is independent of the strength of the stimulus, provided it is above threshold.

When a muscle is in its normal situation in the animal (or when an isolated " nerve-muscle preparation " is used), it is excited by impulses arriving at the nerve-muscle junction in the motor nerve. These release the transmitter substance acetylcholine, which depolarises the end-plate region of each muscle fibre, giving rise to the " end-plate potential " (Chapter 13). In all the striated muscles of the mammals (and in nearly all those of the cold-blooded vertebrates) a single impulse in the motor nerve produces a sufficiently large end-plate potential to initiate a propagated disturbance in the excitable membrane of the muscle fibre. But this is not necessarily so. In most of the muscles of invertebrates (notably those of crustaceans), and in certain " slow " muscle fibres of the frog, the end-plate potential does not set up a propagated disturbance. The muscle fibre then contracts only in the neighbourhood of the end-plates (there are many of them on each fibre) and the magnitude of the contraction varies with the magnitude of the end-plate potential. By using a very small electrode (some 2 μ to 4 μ in diameter) placed very close to a single striated muscle, it is possible to elicit responses only from the part of the fibre adjacent to the electrode. These responses can be " graded " in that as the stimulus is increased a greater length of the fibre is involved ; but before the stimulus has been increased to an intensity sufficient to excite more than a fraction of the fibre the response changes suddenly to a maximal response of the whole fibre. These localised contractions follow localised excitation of the surface membrane. Striated muscle fibres of vertebrates ordinarily contract fully, whenever the stimulus exceeds the threshold, because the contraction is initiated by a wave of excitation, propagated over the whole length of the fibre, which is itself unchanged in size by changing the strength of the stimulus.

Local contractions of a frog's striated muscle fibre are more easily elicited if the small electrode is placed on an isotropic I-band than if it is placed on an anisotropic A-band. It has been suggested that there may be an excitable membrane running transversely across the muscle fibre at the Z-line which runs through the middle of the I-band (see Fig. 16. 13, p. 487) : this carries the excitation rapidly into the interior of the fibre.

Cardiac Muscle. If the auricle of a heart which is at rest is stimulated at any point, and the stimulus is above threshold, the whole sheet of cardiac muscle contracts, and change in the strength of the stimulus does not produce any change in the strength of the contraction. If electrodes are placed on the auricle, diphasic action potentials will be recorded, which are of the same size whatever the strength of the stimulus (above threshold) and wherever the electrodes are placed. This experiment may equally well be done on the ventricle. The whole ventricle will contract, and action potentials, all of the same size, will be picked up from any point. The excitation is normally conducted to the ventricles from the auricles by means of the auriculo-ventricular bundle and the Purkinje tissue which spreads out over the interior surface of the ventricle, as described in Chapter 1, p. 9. These consist of cardiac muscle cells, of rather a specialised kind, but not essentially

different from those in the rest of the heart. Both the excitatory disturbance and the wave of contraction are thus conducted from one muscle cell to another in all directions over the whole heart. This functional continuity of cardiac muscle is a reflection of the protoplasmic continuity of the cells of which it is composed. If any part of the heart contracts, it all contracts (except in abnormal conditions) ; there is no gradation of the contraction with gradation of the excitation, as different fibres are successively brought into action, as there is in skeletal muscle.

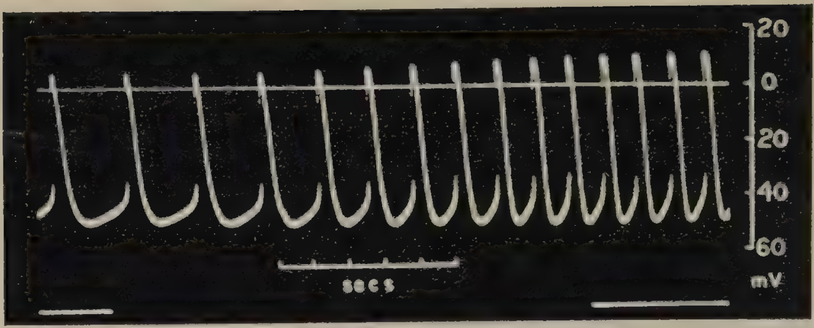
The response of the whole heart is thus " all-or-none," and this type of response was first noticed in heart muscle (by Bowditch in 1871). But it must again be emphasised that many other factors, such as stretching and drugs, can produce changes in the degree of contraction of cardiac muscle, and that variation in intensity of stimulus, with which the all-or-none type of response is alone concerned, may never occur except under experimental conditions.

The spontaneous rhythmic contraction of cardiac muscle is, as we have shown in Chapter 1, an intrinsic property of the muscle and not due to external stimulation. The property is shown in different degree by different portions of the heart, as is demonstrated by the experiments with the Stannius' ligatures described in Chapter 1 (p. 12). The excitable membrane of cardiac cells differs from that of skeletal muscle fibres in that it is not stable in the resting (polarised) condition, but undergoes a relatively slow " spontaneous " depolarisation. When this depolarisation has reached a critical value, there is an " explosive " self-accentuating increase in permeability to sodium ions and potassium ions, successively, just as there is in a nerve fibre or skeletal muscle fibre, and a propagated disturbance is set up. The spontaneous depolarisation in the resting state occurs more rapidly in those cells which constitute the " pacemaker " and in consequence these cells drive the rest of the heart.

Application of adrenaline, or stimulation of the sympathetic nerve fibres supplying the heart, increases the instability ; the depolarisation occurs more rapidly and an action potential is initiated after a smaller delay, as is indicated in Fig. 16. 2 A ; the frequency of the beat is thus increased. Application of acetylcholine or stimulation of the vagus nerve, on the other hand, increases the stability, accelerates the repolarisation of the membrane, shortens the action potential and increases the delay before the next beat is initiated, as indicated in Fig. 16. 2 B. The resting potential, also, may be increased, so that a larger change in membrane potential is needed before the critical value for excitation is reached.

Administration of acetylcholine, or stimulation of the vagus nerve has been shown to increase the permeability of the excitable membrane to potassium ions. In consequence, the membrane potential is more firmly " locked " in a state of maximum polarisation and large depolarising currents are needed before the critical state of depolarisation is reached, at which the " explosive " increase in permeability to sodium ions occurs. Repolarisation of the membrane, after excitation, which results from the delayed increase in

A



B

FIG. 16. 2. Membrane Potentials of Frog's Heart.

Intracellular recording from a " pacemaker " fibre in a spontaneously beating sinus venosus. (The upstrokes of the action potentials, being very rapid, have been lost in the reproduction.)

A. During the break in the lowest reference line, the vago-sympathetic trunk was stimulated at 20/sec., the action of the vagus fibres being inhibited by atropine 1 : 10^6. Note the increase in slope of the slow " pacemaker potential " from which the action potential arises, and the increased overshoot of the action potential.

B. During the four breaks in the reference line the vagus nerve was stimulated at 20/sec. (The tops of the action potentials have been cut off in the recording.) Note the hyperpolarisation, reduction in slope of the pacemaker potential and the rapid repolarisation, as shown particularly in the " escaped beat " at the beginning of the 4th stimulation period. (Hutter and Trautwein, 1956.)

permeability to potassium ions, also occurs more rapidly, as may be seen in Fig. 16. 2 B.

Visceral (Unstriated) Muscle. In most of the mammalian organs which contain unstriated muscle, the muscle cells are in intimate relation with nerve cells and an abundant network of small nerve fibres : changes in the activity of the muscle cells may result indirectly from changes in the activity of these nervous elements, either spontaneous or brought about by any excitatory or inhibitory agents that may be used. It is possible, in the small intestine, for example, to separate the muscular tissue from nearly, if not quite all the nervous tissue, and thus be reasonably certain of studying the properties of the muscle cells alone. Excitation of the nervous elements, moreover, can be distinguished from that of the muscle cells by evidence such as that of the electrical responses, which is highly suggestive, though not always conclusive.

Many kinds of visceral muscle, even after removal from the animal or separation from the organ containing them, are readily excited by *stretching*, or if they are in the walls of a hollow organ, by distension. In most of these muscles there are nerve plexuses which undoubtedly play a part in co-ordinating, and possibly initiating, the activity ; but a bundle of unstriated muscle cells, devoid of detectable nerve cells, may nevertheless respond to stretching and may also develop spontaneous activity, both tonic and rhythmic.

It has been possible to insert micro-electrodes into unstriated muscle cells. There is a resting potential, similar to that in striated muscle fibres, but somewhat smaller and much more variable in magnitude. A feature special to unstriated muscle is that if the particular muscle under investigation shows spontaneous activity, the excitable membrane undergoes transient and periodic partial depolarisation, leading to the appearance of " spikes " on the record of the resting potential ; these may be seen in Fig. 16. 3. If the muscle is stretched, there is a reduction in the membrane potential (the membrane is partially depolarised) and the frequency of the spikes increases. Each spike " takes off " from a temporary and slower reduction of the membrane potential, which suggests that the membrane is inherently unstable and resembles, in this respect, the membrane of cardiac muscle cells which show a " pacemaker " activity. In its response to stretching, on the other hand, the membrane resembles that of a receptor cell which is sensitive to mechanical strain, as discussed in Chapter 17, p. 511. When the muscle is contracting rhythmically, there is a rhythmic decrease and increase in the membrane potential, accompanied by an increase and decrease in the spike frequency ; these fluctuations are " spontaneous " in that they cannot be related to any known influences acting on the muscle.

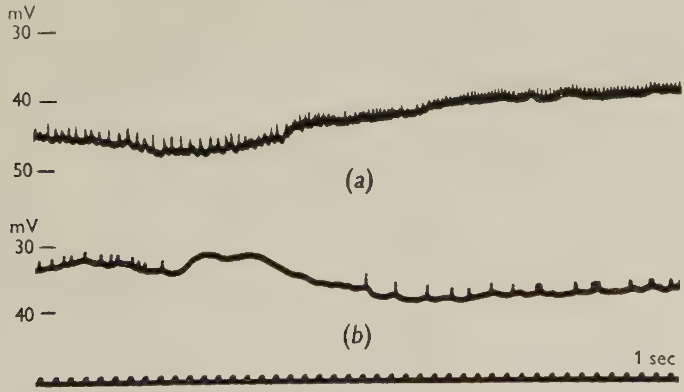


FIG. 16. 3. Membrane Potentials of Unstriated Muscle.

Intracellular recording from fibres of tænia coli muscle of guinea-pig.
(*a*) Depolarisation and increased frequency of spikes produced by acetylcholine (3×10^{-6}) which causes the muscle to contract.
(*b*) Hyperpolarisation and decreased frequency of spikes produced by adrenaline (10^{-8}) which causes the muscle to relax. (Bülbring, 1954.)

If a piece of intestinal muscle is immersed in a very dilute solution of acetylcholine (the excitatory transmitter substance of this kind of muscle) the membrane potential is reduced and the frequency of the spikes increased (Fig. 16. 3 (*a*)). If, on the other hand, it is immersed in adrenaline (the inhibitory transmitter substance) the membrane potential is increased, and the frequency of the spikes diminished (Fig. 16. 3 (*b*)). Direct electrical stimulation of an unstriated muscle fibre which is not spontaneously active, gives rise to a relatively slow wave of depolarisation (analogous to the " local response " of a nerve fibre) on which is superposed, if it is large enough, a transient " spike "; the membrane potential is not necessarily reversed momentarily, as in striated muscle, although it may be.

On repetitive stimulation, at sufficiently short intervals, the slow waves of depolarisation join together, producing a maintained state of depolarisation, like that produced by the action of acetylcholine. But since the spikes become smaller as the depolarisation becomes greater (as a result either of electrical stimulation or of the action of acetylcholine), it is possible to obtain a steady contraction of the muscle which is associated with a maintained, steady condition of partial depolarisation, without detectable " action potentials."

The separate fibres in a bundle or sheet of visceral muscle are not completely insulated from one another, as in skeletal muscle, nor are they completely united, as in cardiac muscle. Excitation of one fibre usually spreads to the neighbouring fibres, but not necessarily over the whole bundle or sheet, and there is no general agreement as to the mode of transmission ; it probably differs from one kind of visceral muscle to another. In some organs, such as the small intestine, there is a progressive wave of excitation which is propagated—or at least co-ordinated—by means of the nervous elements ; but there is no doubt that nerves are not essential for the spread of excitation. An active fibre may stimulate its neighbours either electrically, by means of circulating currents in regions of close contact (compare the transmission of excitation along a nerve fibre), or mechanically, by stretching them.

The Mechanical Properties of Muscles

The response of a muscle to excitation may be made to take a variety of forms. An *isotonic contraction* is one in which the muscle is made to shorten against a constant force. The simplest way of arranging this is to make the muscle lift a constant weight, preferably through a lever whose movement can be recorded, for example, on a revolving smoked drum ; if the weight is hung very near the axis of the lever, its equivalent inertia is reduced and quite rapid shortenings can be faithfully recorded without overshoot or oscillation. An *isometric contraction* is one in which shortening is almost prevented ; the increase of tension is measured in terms of the extension or twisting of a strong spring, and recorded by a light straw writing on a revolving smoked drum, by a mirror which deflects a beam of light directed on to a

rapidly moving photographic plate or paper, or by conversion into electric signals, amplification, and display on a cathode ray oscilloscope. In isometric contraction the muscle performs practically no external work, for work involves movement ; but it does produce energy in the form of heat. In the human body the contractions of muscles are sometimes isotonic, *e.g.* lifting a weight at constant speed, sometimes isometric, *e.g.* holding up a weight. Most often the movement falls into neither experimental type, for the force exerted varies throughout the movement as in pedalling a bicycle, the ejection of blood by the heart, or the discharge of urine from the bladder.

The Latent Period. Following the stimulus to a muscle, there is a short period before any mechanical change can be detected. The apparent duration of this latent period depends on the sensitivity of the apparatus used to record the tension developed—the more sensitive the apparatus the smaller will be the change in tension that can just be detected, and the shorter the latent period. It depends, also, on the time taken for the excitation to spread from the point of stimulation to enough muscle fibres to produce a detectable change in tension. With the most sensitive apparatus so far used, and when the muscle is stimulated directly and more or less uniformly all over, the latent period in a frog's sartorius muscle at 0° C. is only about 8 msec. Curiously enough, the first change in tension is not a rise, but a slight fall—the " latency relaxation "; this is swamped about 15 msec. after the stimulus, by a rapid rise in tension which reaches a value 1,000 times as great as the latency relaxation. At this temperature the whole contraction and relaxation of the muscle last about 800 msec. As the temperature is raised, both the latent period and the duration of the contraction get shorter, being approximately halved for each 10° C. rise in temperature. Roughly, the duration of the latent period is some 1 or 2 per cent. of

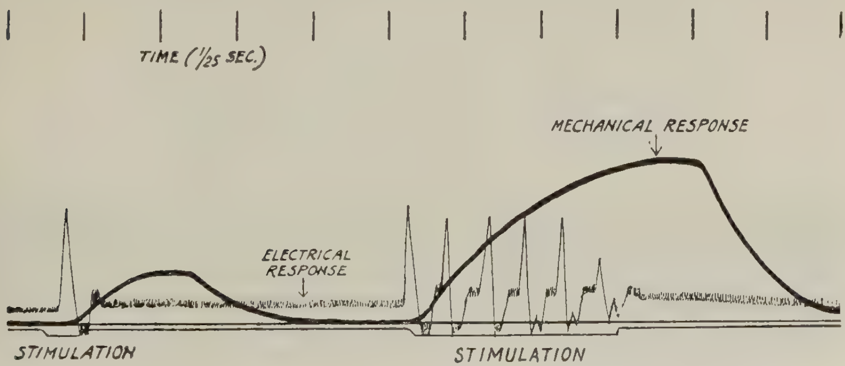


Fig. 16.4. The relation between the **Mechanical and Electrical Responses** of a gastrocnemius muscle of a frog with normal circulation, to (1) a single induction shock, and (2) a rapid succession of six shocks. (These six shocks occurred at times indicated by the corresponding action potentials.) In the twitch the electric response is almost over before the mechanical response begins. In the short tetanus there is a discrete electric response to each shock, whereas the mechanical responses are fused. (After Fulton.)

the duration of the mechanical response, and this figure applies, also, to mammalian skeletal muscle, to cardiac muscle, and probably to visceral muscle ; in this last, the true latent period is difficult to estimate owing to the slow onset of the contraction and the variability in the spread of the excitation.

The Effects of Repeated Stimulation. If a single adequate stimulus, such as a brief pulse of electric current, is applied to a skeletal muscle or to the motor nerve supplying it, the muscle responds by a brief contraction known as a " twitch." If two such stimuli are applied in succession, the result depends on the interval of time between them. If the interval is short enough, the second stimulus has no effect whatever ; muscle fibres, like nerve fibres, have an *absolute refractory period*, during which they are completely inexcitable. This is followed by the *relative refractory period*, during which they will respond if the strength of the stimulus is increased sufficiently. The duration of the refractory condition depends on the kind of muscle studied and, as in nerve, is related to the duration of the action potential—*i.e.* to the period of excitation and recovery of the surface membrane. Its significance, however, depends not so much on its absolute magnitude as on its relation to the time taken by the muscle to contract and relax again.

In *skeletal* and *visceral muscles* the depolarisation and recovery of the excitable membrane, and thus the refractory period, are over well before the development of tension, or shortening, of the contractile elements of the muscle has been completed, as is shown in Fig. 16. 4. A second stimulus, applied shortly after the end of the refractory period will thus produce a contraction which is superposed on the contraction set up by the first stimulus, and the two contractions are said to *summate*. A succession of stimuli will produce a more or less sustained contraction, since the muscle will not have time to relax completely before it is made to contract again by the stimulus, as shown in Fig. 16. 5. If the interval between the stimuli is so small that no relaxation can be detected, the prolonged mechanical response is called a *tetanus*, evoked by *tetanic stimulation*. The least frequency of stimu-

FIG. 16. 5. Records showing progressive **Fusion of Individual Mechanical Responses** of a muscle, as the frequency of stimulation is increased. Note that the individual responses last longer after a preceding period of activity, so that a sustained contraction may ultimately be maintained at an economical frequency of stimulation, which is low enough to allow intermittent relaxation when first applied. (From Howell, " Textbook of Physiology.")

lation required to produce a tetanus will depend on the type of muscle, the temperature, and the sensitivity of the apparatus used to examine the smoothness of the tension record. In frog's muscle at 0° C., the fusion frequency is about 25 per second ; in the internal rectus of the cat's eye at 37° C. (a very fast muscle), about 250 per second. In visceral muscle, on the other hand, owing to the very slow relaxation (exceedingly slow in some kinds of muscle) frequencies of less than 1 per second, or even 1 per minute, may be sufficient.

FIG. 16. 6. Tracings of **spontaneous contractions of frog's ventricle,** to show **refractory period** and **compensatory** pause. (Marey.)

In *cardiac muscle* the action potential lasts very much longer than it does in skeletal muscle, and the membrane does not recover until the contraction is almost completed. Correspondingly, the refractory period lasts almost as long as the whole period of contraction and relaxation. If two stimuli are applied in succession, the second stimulus produces no effect at all unless it is applied so long after the first that the responses of the two are independent contractions occurring one after the other. *No summation* of contractions and *no tetanus* can be produced under normal conditions in heart muscle. If the second

stimulus occurs during the relative refractory period, the contraction produced is usually smaller than a normal one, owing to imperfect recovery of the contractile process, as may be seen in Fig. 16. 6.

Let us observe a spontaneously beating heart, and interrupt its rhythm by interpolating an electric shock applied to the ventricle. Fig. 16. 6 shows the ventricular contraction during such an experiment, in which a series of shocks was applied, each at a slightly later phase of the normal contraction. In the lower three records, the stimulus fell in the absolute refractory period and produced no effect. In the upper ones it fell later and produced a smaller or greater contraction according as it occurred sooner or later. This contraction took place, abnormally, before its time, and was accompanied by its own refractory period. It was followed by a pause known as the *compensatory pause*, before the next (normal) contraction, because one of the regular stimuli transmitted from the pacemaker found the ventricle in a refractory state and so produced no effect. Comparable effects are sometimes produced in the human heart by abnormal contractions (premature contraction or extra-systole) originating at some irritable focus independently of the normal rhythm.

If the frequency of the stimuli initiated by the pacemaker (or applied artificially) is progressively increased, a time will come when the interval between successive stimuli is so short that it is less than the refractory period of, say, the ventricular muscle ; the muscle will then suddenly respond only to alternate stimuli. If the frequency be further increased so that of three stimuli, the second and third fall within the refractory period accompanying the response to the first, the muscle will respond only to every third stimulus. Analogous events are observed in certain cases of " heart-block " in human disease as discussed in Chapter 1, in which the ventricle responds to only a fraction of an abnormally rapid sequence of stimuli coming from the auricle. If the frequency of stimulation be constant, and the duration of the refractory period be sufficiently increased, a similar omission of the response to alternate stimuli may be observed.

If the rate of stimulation be increased gradually, the refractory period will itself become gradually shorter and so enable the heart to beat faster than it could have done had the rate of stimulation been suddenly increased. This may explain the fact that in certain conditions the human heart can attain a rate of 200 beats per minute ; such a rate would be impossible without an actual shortening of the normal refractory period.

Voluntary Contractions. In the intact animal or man, most skeletal movements require that the muscles concerned should exert forces for an appreciable period of time. Although a few movements, *e.g.* blinking the eye-lids, occur so rapidly that they might be produced by muscular twitches, the vast majority must be produced by tetanic stimulation. Many of these involve only submaximal contractions and this seems to be brought about by tetanising different parts of the muscle in rotation. In this way each group of fibres is regularly rested, so that small tensions

can be maintained for a long period without fatigue of the muscle as a whole.

The electrical changes which accompany muscular activity have been extensively studied in a large number of human movements, and records of such changes (which are called electromyograms or EMG's)

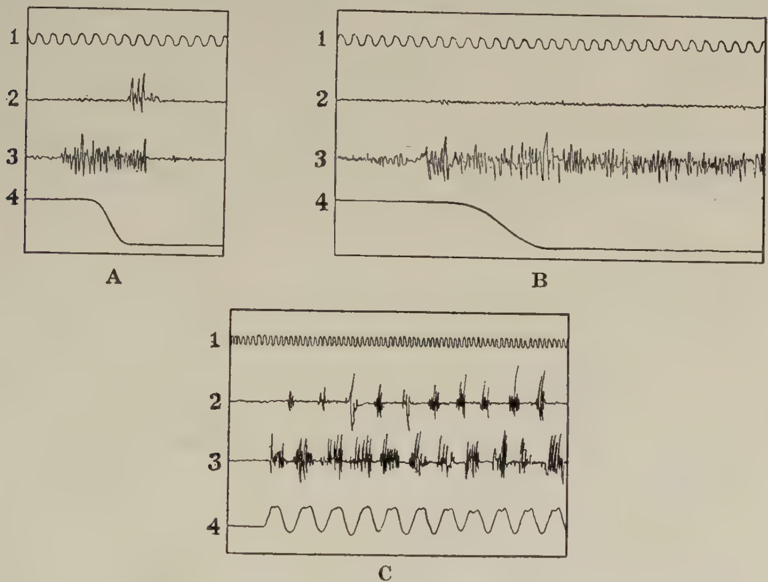


FIG. 16. 7. Electromyograms from Human Muscles during Maximal Voluntary Flexion of the Elbow.

Each record shows from above downwards : (1) time marker, 20 cycles per second ; (2) action potentials in triceps muscle ; (3) action potentials in biceps muscle ; (4) movement of hand, downward deflexion indicating flexion of arm.

A. No external load. Movement is checked by contraction of triceps.

B. Movement against a force of 5·5 kg. wt. Movement is slower and is checked by the external force.

C. Rapid to-and-fro movement of arm showing alternate activity in biceps and triceps. (D. R. Wilkie.)

have been used clinically to assess muscular function. The exact type of record obtained depends on the type of recording electrodes used and on the degree of activity in the muscle. If the record is made from a very small volume of muscle it is possible to distinguish a succession of sharp spikes whose *frequency* rises as the activity increases in the neighbouring muscle. Such records are easily obtained from a pair of electrodes arranged concentrically—*e.g.* a hypodermic needle containing an insulated wire with a bare tip—which is sensitive only to potential changes in the region of the tip. However, if the recording is made from a pair of electrodes spaced an appreciable distance apart the character of the record is quite different. The individual responses from different parts of the muscle are then partly added, partly cancelled out, and the result is a complex, irregular wave whose *amplitude* increases as activity

increases. Electromygrams such as those shown in Fig. 16. 7, can be used to indicate when a muscle is active, and to indicate roughly its degree of activity. They show clearly that even a brief movement involves tetanic stimulation.

The tension developed in an isometric tetanus is surprisingly large, *e.g.* 1,000 g. wt. by a frog's gastrocnemius weighing 0·5 g., and 50 kg. wt. by the gastrocnemius of a cat. There has been a good deal of discussion as to whether such large tensions are ever developed during life. It is not normally possible to decide the question by comparing a maximal voluntary effort with an artificial tetanus in the same muscle, though this has been done successfully with the adductor of the thumb, where it is found that the tetanic tension and the voluntary tension are

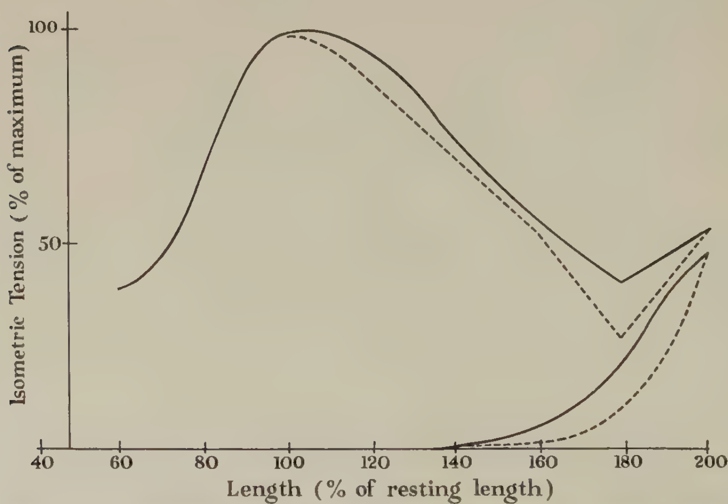


FIG. 16. 8. Isometric tension-length curve of tetanised single muscle fibres. The active tension is shown in the upper curves, and the resting tension in the lower curves. The full line passes through points obtained with increasing tensions ; the dotted line through those obtained with decreasing tensions. (After Ramsey and Street.)

the same. The arguments against their being the same in other muscles rest on very indirect evidence, such as the occurrence of athletic injuries in which damage is done to tendons or bony attachments, supposedly as the result of a stronger-than-usual contraction. However, most of these injuries seem to be the result not of a smooth contraction, but of a sudden jerk in which energy stored in other moving parts of the body is suddenly dissipated in the affected region.

The Tension-length Curve. Resting muscle is an elastic body, which means that when a tension is applied, it stretches to a new equilibrium length. In a whole muscle this property depends to some degree on the presence of connective tissue binding the fibres together, but even isolated single fibres are elastic, as shown in the lower curves of Fig. 16. 8.

The active tension developed on tetanising a muscle or muscle fibre also depends on the length at which the stimulation took place (upper curves of Fig. 16. 8, Fig. 16. 9). The length of a striated muscle in the body, which is decided by the arrangement of the skeleton, is normally found to be that which permits the greatest tension development. This is not true of cardiac muscle, which normally works only on the left-hand (rising) side of its tension-length curve. This means that

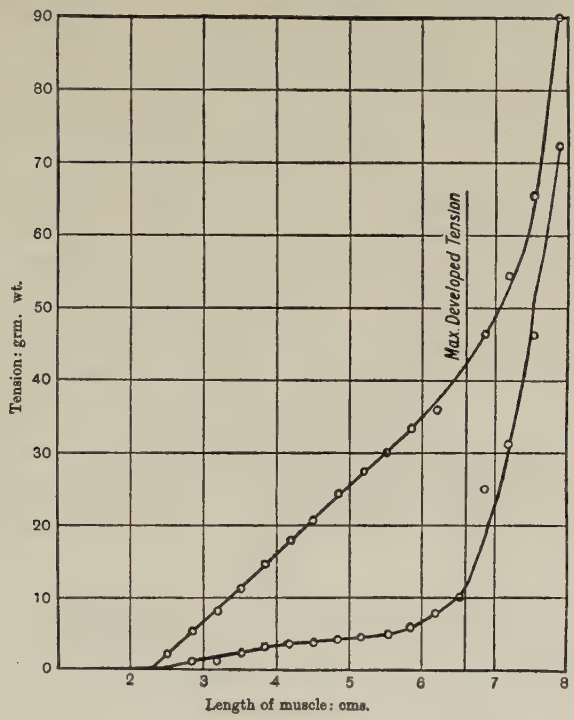


Fig. 16. 9. Tension-length Curve of Unstriated Muscle (dog's retractor penis). The upper curve shows the tension of the electrically stimulated muscle, the lower curve that of the unstimulated muscle. Note the increase of tension development (difference between the two curves) with increase of length, up to an optimum length, as in striated muscle.

increased diastolic filling stretches the muscle fibres to a length where they are able to develop more tension, as described in Starling's law of the heart, and illustrated in Fig. 1. 12 (p. 22). Many kinds of visceral muscle have spontaneous activity, and no precise " resting " length in the absence of external stimulation : the isolated retractor penis of the dog, however, a small strip of muscle, free from nerve cells and with fibres running parallel to one another, is not spontaneously active, and the tension-length curve is well defined (Fig. 16. 9).

Skeletal muscles normally work against substantial loads, and do not change very much in length—from being stretched to about 110 to 120

per cent. of the " resting " length, they shorten to about 80 or 90 per cent. of this length (the figures vary considerably according to the function and situation of the muscle). Visceral muscles, on the other hand, usually work against relatively small loads and many of them have been observed to shorten to one-quarter, or less, of the extended length.

The Force-velocity Curve. The most important function of muscles is to produce mechanical work, and the simplest way to study this production of work is in isotonic contractions. By letting the muscle shorten against various isotonic loads, it has been found that the *speed*

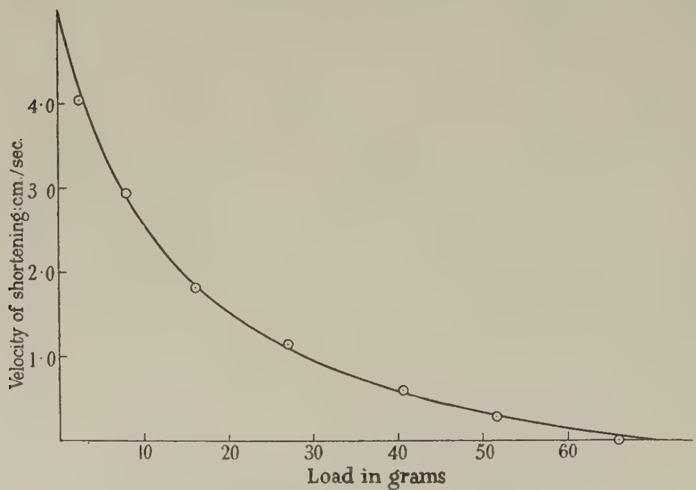


FIG. 16. 10. Relation between load (g. wt.) and speed of shortening (cm./sec.) in isotonic contraction. Each circle is the mean of two observations in a series and reverse.

Muscle: Frog's sartorius, 165 mg., 38 mm., in Ringer's solution at 0° C. Tetanus 11·4 shocks/sec. Observations at 4-min. intervals: 5·6 mm. shortening.

The curve is calculated from the equation $(P + 14{\cdot}35)\ (v + 1{\cdot}03) = 87{\cdot}6$. (A. V. Hill.)

of shortening depends in a predictable way on the applied force—the greater the force, the smaller the speed and *vice versa*. This relation is illustrated in Fig. 16. 7 (compare line 4 in records A and B), and is familiar to us in everyday life, for we know from experience that we cannot lift a heavy weight as quickly as a light one. The exact relationship between force and velocity of shortening is shown in Fig. 16. 10 for a frog's sartorius muscle at 0° C. This figure also demonstrates A. V. Hill's discovery that experimental force-velocity curves can be accurately fitted by the equation :

$$(P + a)(v + b) = \text{constant},$$

where P is the isotonic force, v the velocity of shortening and a and b are constants. Hill's equation has been found to describe accurately

the behaviour of all the muscles so far tested, including those from the frog, man, ray, snail, mussel and tortoise ; in the last three animals the muscles used were unstriated. There is no reason to suppose that it does not apply also to cardiac muscle, although this has not actually been demonstrated. Moreover, all these muscles are found to exert about the same tension (reckoned per unit cross-section) ; the chief difference between them is an enormous variation in their speed of shortening, which is in every case adapted for the function which the muscle is to perform in the body.

It should be noted that since the total amount of shortening is limited by the tension-length curve, the velocity of shortening against a given constant force must diminish as shortening proceeds. However, the muscle continues to obey Hill's equation (in a slightly modified form) throughout its whole contraction.

The force-velocity curve is of great practical importance in the economy of the body as a whole, for the rate at which work is done ($= P.v$) clearly must depend on the load, being small for very large

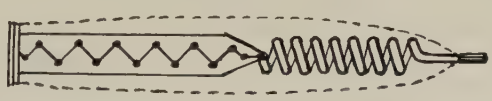


FIG. 16. 11. Diagram showing how a complete muscle (indicated by the broken lines) behaves as if it consisted of a contractile component joined to the end of an inert elastic component.

loads and for very small ones, and passing through a maximum somewhere in between. The mechanical leverages of the skeleton are so arranged that the muscle will, in so far as possible, be presented with optimal conditions of load and speed. The same principle applies in the design of gearing for bicycles and in the design of many tools. In passing it may be noted that the maximum rate of working is surprisingly high—for example, maximal flexion of the elbow can generate one-fifth of a horsepower. In favourable conditions, as in rowing, a strong man has been observed to exert continuously and for some time, nearly $\frac{1}{2}$ h.p.

The Contractile and Elastic Components of Muscles. The mechanical properties of muscles may be conveniently described in terms of a "model" which consists of two separate components, as indicated in Fig. 16. 11. One of these, the contractile component, shortens when it is activated by appropriate stimulation ; if the ends of the muscle are fixed, or attached to a load, it thus stretches the other, elastic, component and sets up a tension in it.

The contractile component always shortens to the same final length, whatever its initial length (within limits) ; thus if the contraction is isometric, the greater is the initial length, the more it extends the elastic component and the greater is the tension set up. Obviously, if the

muscle is allowed to shorten freely until the contractile component is fully shortened, before being called upon to exert tension, then no tension can be set up. If, however, the muscle is stretched initially to such an extent as to produce a tension even in the resting state, the contractile element may not be strong enough to shorten completely against the pull of the elastic component ; the total tension continues to rise, although the extra tension developed on stimulation begins to fall (compare Figs. 16. 8 and 16. 9).

On stimulation, the contractile component goes into what has been called the " active state " almost immediately and shortens at a speed which depends on the load opposing it. After each stimulus, the active state is maintained at its full level for a short time and then disappears gradually, so that the muscle relaxes. If the ends of the whole muscle are held rigidly, and the contraction is isometric, the contractile component shortens relatively slowly, since it is opposed by the force exerted by the elastic component. This is why the isometric tension rises relatively slowly to its maximum value, in spite of the almost instantaneous " activation " of the contractile component. In a twitch, following a single stimulus, the active state does not last long enough for the contractile component to shorten completely ; if there is a succession of stimuli, as in a tetanus, the repeated renewal of the active state gives the contractile component time to shorten further and to stretch the elastic component more completely. The isometric tension developed in a tetanus is thus greater than that exerted in a single twitch, as is demonstrated in Fig. 16. 4.

If the muscle is given an appropriate stretch, immediately after the end of the latent period, it immediately develops the maximum tension which is reached more slowly in a tetanus : the elastic component is extended by the stretch and the contractile component does not have to shorten. This is one of the lines of evidence showing that the muscle becomes fully active almost instantaneously ; the other is derived from the rate of heat production, to be discussed in the next section.

Many kinds of unstriated muscle give no detectable response to a single stimulus, but a substantial response if the same stimulus is repeated a sufficient number of times, the number becoming smaller as the intensity of the stimulus is made greater. This may be due only to the fact that the active state lasts for a very short time compared with the time taken for the contractile component to shorten—*i.e.* that there is a summation of contractions, each too small to be detected by itself. But it is possible, also, particularly in those kinds of visceral muscle which are associated with nerve cells, that there is a temporal summation of sub-threshold stimuli, as is known to occur, for example, in the central nervous system.

Heat Production by Muscles

Everyone knows that muscular work is accompanied by the production of heat. Muscles are not perfectly efficient, and a considerable fraction (about 4/5) of the chemical energy used up when they do

mechanical work is " wasted " as heat. This circumstance has proved to be of great value in studying the nature of muscular contraction, since in the hands of A. V. Hill and his colleagues, it has been found possible to measure the changes in temperature of an isolated muscle, and to analyse the time relations of the processes responsible for them, with considerable accuracy.

The changes in temperature amount to only a few thousandths of a degree, and much of the heat which gives rise to them is produced during the contraction, which lasts only a fraction of a second, at rates which vary from moment to moment. Nevertheless, sufficiently delicate and elaborate techniques have been evolved, and it is possible to observe and measure the changes in temperature of a muscle every few milliseconds, even though they amount to only a few millionths of a degree. Nearly all the observations have been made on sartorius muscles of frogs and toads at 0° C.

As in all kinds of living tissue, metabolic changes continue in muscles even when they are at rest, so long as the cells are alive. There is, accordingly, a steady production of *resting heat* (about 20–30 microcalories per minute in a frog's sartorius muscle at 18° C.—enough to raise its temperature by about 1/100 of a degree per hour if no heat were lost to the surroundings). If a muscle is given a single maximal stimulus and responds with a twitch, the rate of heat production increases greatly, about a hundred-fold or more, above its resting value, and remains increased for a considerable time after the muscle has relaxed again. That part of the total quantity of extra heat produced which appears during the contraction and relaxation of the muscle is known as the *initial heat* ; that part which appears subsequently is known as the *recovery heat.*

(1) The *Initial Heat* is made up of three components, which may vary independently:

(*a*) Immediately the muscle begins to contract *activation* (or *maintenance*) heat is produced. This starts abruptly and continues at a constant rate so long as the muscle is kept fully active (by tetanic stimulation, for example) ; the active state is set up, and maintained, only by the continuous expenditure of energy. If the contraction is isometric and no external work is done, very nearly all the heat produced during the contraction is activation heat (Line AA in Fig. 16. 12).

(*b*) If the muscle is allowed to shorten and do mechanical work, extra heat (the *shortening* heat) appears (line AB in Fig. 16. 12) ; the magnitude of the shortening heat is proportional to the *distance* through which the muscle has shortened, and is independent of the load.

(*c*) During relaxation, if the muscle is lengthened by the load, *relaxation* heat appears (line BC in Fig. 16. 12). This is the heat, dissipated in the muscle, which is equivalent to the mechanical work expended in raising the load. If the apparatus is so arranged that the load is detached from the muscle at the height of the contraction, it retains the potential energy given to it, and no relaxation heat appears in the muscle (line BB in Fig. 16. 12). The small relaxation heat under isometric conditions represents energy stored in the elastic component of the muscle which is degraded into heat during relaxation.

(2) *Recovery Heat* is produced in the muscle for 20 to 30 minutes after the end of the period of activity (in frog's skeletal muscles). The total amount produced is just about equal to the total amount of initial heat produced ; when the muscle is allowed to shorten and do work, it produces extra recovery heat as well as extra initial heat. If the muscle is

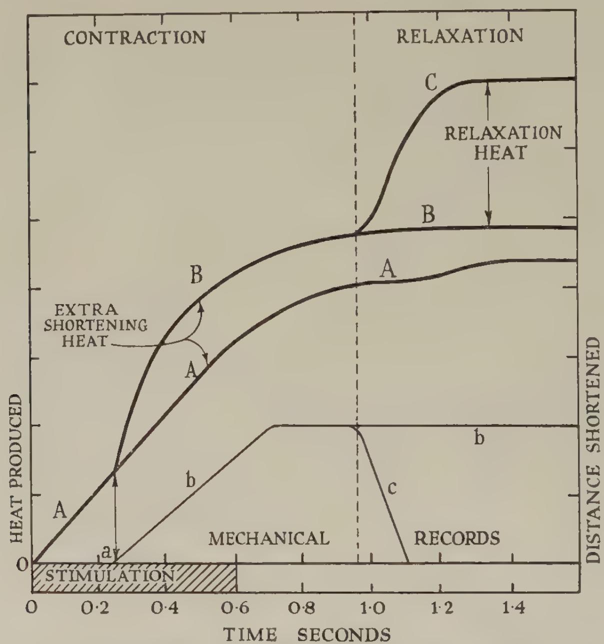


FIG. 16. 12. Diagram showing the relation between **Heat Production** and **Mechanical Work** in an isolated muscle.

Heavy lines—amount of heat produced.
Thin lines—changes in length of the muscle.
Repetitive stimulation is maintained for 0·6 seconds.
AA.—isometric contraction. No mechanical work done.
ABB.—isometric up to the arrow, then isotonic. Weight held up during relaxation ; no relaxation heat.
bb.—corresponding mechanical record.
ABC.—exactly as ABB but the muscle is lengthened by the load on relaxation. The work done by the load on the muscle is converted into heat.
bc.—corresponding mechanical record.

(After A. V. Hill.)

not provided with oxygen, *i.e.* in anaerobic conditions, the amount of heat produced during recovery is only about one-fifth to one-twentieth of that produced in the presence of oxygen ; it must, therefore, be produced mainly by oxidative chemical reactions. The nature of the various reactions concerned will be discussed in a later section. Muscular activity, therefore, is like that of a group of primitive warriors who, from the top of a precipice, drop a huge stone on their enemies ; having

done so, they laboriously roll another stone up the hill in preparation for the next engagement.

Thus in summary, if a muscle, supplied with oxygen, is stimulated and lifts a weight P through a distance x, the amount of initial heat will be given by :

$$H = A + a.x + P.x$$

where A is the activation heat, $a.x$ is the shortening heat and $P.x$ is the energy imparted to the weight ; if the weight is lowered again, stretching the muscle during relaxation, $P.x$ is also the relaxation heat. The quantity H is approximately equal to the recovery heat so that the total energy used by the muscle is approximately $2\,H$.

Muscular Efficiency. Optimum Speed of Movement

By measuring the amount of mechanical work done, and either the total amount of heat produced or the amount of chemical energy released (deduced from the amount of oxygen consumed), it is possible to calculate the efficiency of the contractile process, *i.e.* the amount of mechanical work produced from unit quantity of chemical energy. As pointed out in an earlier section (p. 489), the inverse relation between the force exerted by a muscle and the speed at which it contracts, implies that there will be an optimum speed at which the greatest amount of work can be done. But the amount of energy released does not depend on the speed of the contraction ; activation heat is produced even if there is no shortening and no work is done ; any additional shortening heat depends only on the amount of the shortening and not on its speed. The efficiency of the contraction, therefore, will also have a maximum value, and will approach zero when the shortening is either very slow or very fast. This maximum efficiency is about 20 per cent, both for frog's isolated muscles (from measurements of heat production) and for human movements (from measurements of oxygen consumption).

That extreme slowness or quickness of movement is extravagant of energy is familiar in most human activities ; a man lifting a weight almost too heavy for him works hard but gains little ; a fast bowler endows the ball with a kinetic energy which is but a small fraction of the energy that he uses to deliver it. In going upstairs, one of A. V. Hill's colleagues discovered, by actual measurement, that about one step per second was his most economical speed, and that more oxygen was used for the same amount of work if he went either more quickly or more slowly. In marching, a given distance can be covered most economically at a rate of about 100 paces a minute. The optimum speed, however, is usually not well defined ; there is little change in the efficiency in this region over a range of speeds of nearly three-fold. Running, on the other hand, is a form of exercise in which there is no optimum speed, since it is not possible to perform the movements of running very slowly. The energy required to run a given distance varies approximately as the cube of the speed, and that required to run for a given time as nearly the fourth power of the speed.

Muscular Tone

In physiology, the word " tone " (or " tonus "), as applied to a muscle, defines its ability to develop or resist a force for a long period of time without change in length, or to change in length while maintaining a constant tension. But the word has been used, both in physiology and in medicine, with rather a wide variety of meanings. For example, it has been used clinically to describe the condition of an organ, *e.g.* the " tone of the skin," even " mental tone," both suggesting vaguely a criterion of resiliency. The pharmacist's " tonic " implies a similar meaning. A heart with good " tone " is one which is in good condition and is therefore not dilated for, as described in Chapter 1, p. 28, dilatation is one of the signs that the heart is calling up its functional reserves. On the same lines an " atonic " bladder or uterus is one which is dilated and contracts poorly. Such uses of the word are dangerous, since the meaning is vague and ill-defined.

When a muscle is held in a steady contracted condition, no external work is done, but chemical energy is released and heat produced. The smaller is the rate of heat production, for a given tension in the muscle, the more " efficient " is the muscle in maintaining such a *tonic* contraction. (The word " efficient," of course, is not being used in its strict sense, as defined in the previous section.) The less frequently the muscle fibres are, on the average, stimulated, the less heat will be produced in a given time. The more slowly a muscle relaxes, the less frequently need it be stimulated in order to maintain a " fused " tetanus, and thus the more " efficient " it is in maintaining a tonic contraction.

Unstriated muscles, owing to their extreme sluggishness, particularly in relaxation, are well adapted for maintaining tonic contractions. Indeed, some of them are able to remain contracted for long periods of time without obvious repetitive excitation and with no apparent expenditure of energy. This circumstance led to the idea that some kinds of unstriated muscle, *e.g.* those that hold the shells of the common mussel closed, possess a " catch " mechanism, analogous to a ratchet device, by which they can maintain a tension passively, without continuous excitation. There is little evidence to support this theory ; the necessary frequency of tetanic excitation is very small, and it can be calculated that energy will be expended so slowly that it would very probably escape detection ; moreover, as mentioned on p. 480 above, there may be maintained states of depolarisation, without detectable " spikes."

The sluggishness of unstriated muscles may be observed when such a muscle is stretched by a constant weight under resting conditions ; it may not reach its new equilibrium length for anything up to an hour or more and behaves as if it were more " viscous " than striated muscle. This " viscosity " is probably a property of the contractile component of the muscle, for it is greatly affected by the presence of certain drugs (*e.g.* adrenaline greatly reduces the resting " viscosity "

of the dog's retractor penis muscle) and by certain types of previous stimulation. The molluscan muscles which hold the shells together are provided with inhibitory nerves which greatly increase the rate of relaxation of the muscle—an effect which is very like that of releasing a " catch " or ratchet.

Mammalian skeletal muscles are sometimes said to be in tonic contraction when maintaining posture, or in " decerebrate rigidity " (Chapter 14, p. 415). There is little expenditure of energy and the electrical changes that would be expected if there were tetanic excitation cannot always be detected. It is not necessary, however, to suppose that there is a " catch " mechanism. These contractions are regulated by stretch reflexes (Chapter 14, p. 408) ; if, in a decerebrate preparation, the experimenter stretches a muscle slightly to see whether it is tonically contracted or not, the contraction promptly appears ; when the stretch is removed, the contraction vanishes. Even during the maintenance of posture by a normal animal or man, many of the muscles are not continuously active. When a man is standing, for example, the leg muscles contract only when the body becomes slightly off-balance, and is then returned to its central balanced position ; when this happens, the electrical signs of tetanic excitation can always be observed. When a skeletal muscle is voluntarily relaxed, no electrical changes can be detected and there is every reason to suppose that its tone disappears completely.

The Contractile Engine

Muscles of all kinds are essentially engines in which energy released in chemical reactions is converted into mechanical work by which a load is moved. In describing how the engine works (in so far as this is known) we have to consider the nature of the reactions which supply the energy, and the nature of the structures, or " machinery," which are responsible for the shortening and the development of tension.

The Contractile " Machinery." Like all living cells, muscle fibres and muscle cells contain substantial amounts of proteins of various kinds. But among these there is a special kind, of particular importance, known as *actomyosin*, which does not occur in other kinds of cell ; it is a complex of two proteins, *actin* and *myosin*, which may be obtained from minced muscles by careful extraction with salt solutions. There is strong evidence that actomyosin is the essential part of the contractile " machinery " and is responsible for the shortening and development of tension by the muscle. There are two ways in which it may do this ; they are not mutually exclusive and it is not yet generally agreed which is the more important.

(1) All proteins consist of amino-acids linked by peptide bonds into long chains, like beads strung on a necklace. The valency bonds of each carbon atom stick out at fixed angles (they point towards the corners of a triangular pyramid), and the chain can never be pulled out perfectly straight, but only into a helical (corkscrew) form. Those parts of the amino-acids other than the $-NH_2$ and $-COOH$ groups

(which unite to form the peptide bonds) stick out from the main corkscrew as " side-chains "; these may be quite long and elaborate structures. Now although the bond angle is fixed, it is easy for one part of the molecule to swivel, relative to another part, around a bond as axis. In this way the main chain may fold up and become shorter, each turn of the corkscrew becoming larger and containing many more amino-acids than before. Such a folding is likely to occur, for example, if there are strong attractions between the side-chains.

In the " fibrous " proteins, such as the keratin of hair, the main chains are associated in bundles, like a handful of match-sticks, some of the side-chains of each molecule being " cross-linked " to some of those of its neighbouring molecules. All the main chains are thus held in a more or less constant state of folding, and the whole bundle resists stretching. But if the cross-links are broken by some appropriate treatment, the main chains become free to fold up, or be stretched out. The distance between the regularly occurring side-chains of a protein molecule may be measured by means of X-ray diffraction analysis, and it has been shown, for example, that the shrinking of wool fibres when exposed to hot water is due to an increased folding of the keratin molecules. Actomyosin is also a fibrous protein. It is tempting to suppose that muscular contraction results, similarly, from an alteration in some of the side-chains, leading to a change in their mutual attractions and repulsions, and thus to an increased folding of the actomyosin molecules. But so far, no definite evidence has been obtained from X-ray diffraction studies that this does in fact occur.

(2) By treating muscles with salt solutions of appropriate concentration, it is possible to extract the myosin only, leaving the actin, and then, by further treatment, to extract the actin. When striated muscles are thus treated and examined in an electron microscope before and after extraction, it is seen that myosin occurs only in the A-(anisotropic, dark)-bands, and is present in the form of little rods. Actin, on the other hand, occurs in the I-(isotropic, light)-bands and in the outer parts of the A-bands but not in the central parts (H-zones), and is present as fine filaments, about one-third the diameter of the myosin rodlets. If the muscle is viewed in cross-section, it is seen that except in the H-zones, each rodlet is surrounded by an array of filaments. When viewed in longitudinal section (Fig. 16. 13 (*a*)), the actin filaments can be seen, in the A-bands, to run in between the myosin rodlets, stopping short at the edge of the H-zones. When the muscle is lengthened by stretching, or allowed to shorten by contraction (within physiological limits) the width of the I-bands changes correspondingly, but the width of the A-bands remains unchanged. As the I-bands become narrower, for example, the muscle becoming shorter, so do the H-zones, and, as may be seen in Fig. 16. 13 (*b*), it seems that the actin filaments are drawn further in among the myosin rodlets, pulling the Z-lines after them. Conversely, on stretching the muscle, the actin filaments seem to be pulled out from among the myosin rodlets ; both the I-bands (without rodlets) and the H-zones (without filaments) thus become wider.

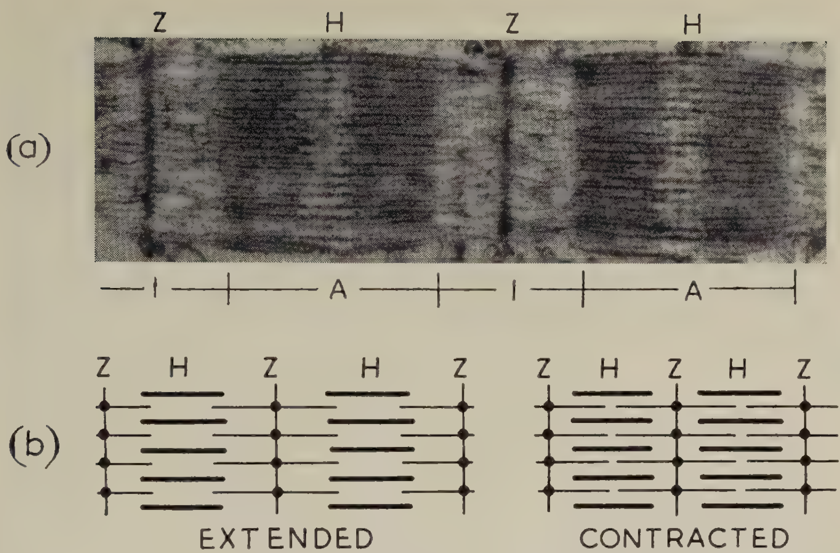


FIG. 16. 13. The Fine Structure of Skeletal Muscle.

(*a*) Electron micrograph of very thin longitudinal section of rabbit's psoas muscle, extracted with glycerol and fixed with buffered osmium tetroxide. The thicker rodlets (of myosin) are seen in the A bands only : the thinner filaments (of actin) are seen in the I bands and running in between the rodlets in the A bands as far as the H zone. The length of each A band is about 1·5 μ. (H. E. Huxley.)

(*b*) Diagram showing how the thin actin filaments appear to slide in and out between the thicker myosin rodlets as the muscle changes in length. When the muscle is extended by stretching, the length of each sarcomere (*e.g.* from one Z line to the next) is about 3 μ; when the muscle is contracted the sarcomere length is about 2 μ. (After Hanson and Huxley.)

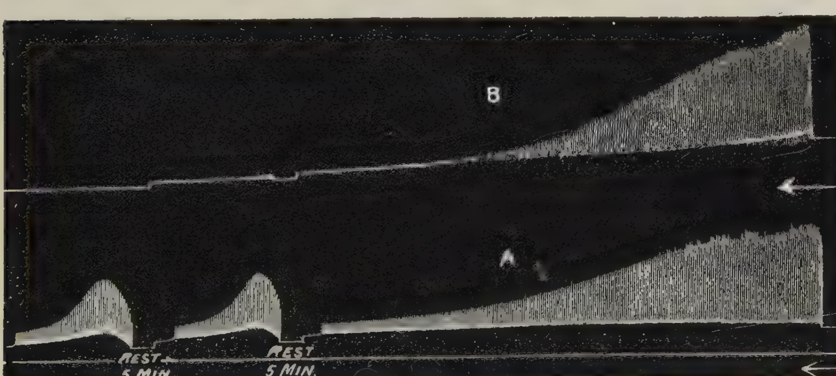


FIG. 16. 14. Showing that oxygen is required for recovery, not for contraction.

Sartorius muscles A and B were both stimulated once per second and allowed to contract isotonically under a load of 6 g. (temp. 19° C.). Muscle A was exposed throughout to oxygen ; muscle B was in nitrogen. Notice that the time scale runs from right to left. (Fletcher.)

It is suggested, therefore, that in striated muscle the " machinery " consists of this system of sliding filaments. On activation of the muscle some changes occur on the side-chains of the actin and myosin molecules and an attractive force is developed between them such that the filaments are drawn in among the rodlets, just as an iron core is drawn into a solenoid when current is passed through the windings. Just how this could be brought about by means of inter-molecular forces between the side-chains is still somewhat obscure.

TABLE 16. 1.

The Approximate Concentrations of Certain Substances in Frog Skeletal Muscle (expressed as a percentage of the weight of the tissue).

	Living		Dead
	At rest	After contracting (fatigued to exhaustion)	(in rigor mortis)
Adenosine triphosphate*	0·5	0·5	0
Creatine phosphate* . .	0·5	0·1	0
Glycogen . . .	0·7	0·5	trace
Lactate* . . .	trace	0·2	0·8

* Present mainly as potassium salt.

The Supply of Energy. It has been known for a long time that isolated muscles, like intact animals, absorb oxygen and give off carbon dioxide, and that these processes are accelerated by activity. In the last resort, muscles depend on oxidative reactions for the supply of energy ; but the observations on the time relations of the heat production discussed earlier in this chapter, and experiments such as that illustrated in Fig. 16. 14, show that the *immediate* source of energy for muscular contraction is a reaction which does not involve oxygen. Muscle B, in Fig. 16. 14, which was in nitrogen, contracts initially as well as does muscle A which was in oxygen. That the latter " tires " nearly as fast as the former is due to the fact that oxygen cannot diffuse into it sufficiently rapidly, and most of the fibres became anoxic. But when both exhausted muscles are given a rest, the one in oxygen recovers, while the one in nitrogen does not.

In order to discover the nature of the non-oxidative reactions, we must know what substances are available within the muscle fibres, and whether their concentrations change when the muscle contracts. Of all the various substances present, the relevant ones, and their concentrations, are listed in Table 16. 1.

Adenosine triphosphate (ATP) is a substance which is known to play an important part in many kinds of metabolic reaction. It is one of the " active " (" energy rich " or " high energy ") phosphate compounds, and by losing one of its phosphate groups, becoming adenosine

diphosphate (ADP), it can supply energy and drive chemical reactions which would not proceed by themselves (Chapter 6, p. 204). That it probably supplies energy to the contractile machinery of a muscle is suggested by two lines of evidence. First, the protein myosin, as extracted from muscles, contains prosthetic groups which catalyse the hydrolysis of ATP to ADP, thus releasing energy. Secondly, filaments of actomyosin will shorten and do work when they are treated with ATP. Such filaments can be made by extruding actomyosin solutions through a fine jet, or more simply, by washing living muscle fibres with dilute glycerol which breaks down the cell membranes and allows many of the muscle constituents to diffuse away, leaving the isolated protein skeleton *in situ*. In either case the actomyosin is inextensible until ATP is added, whereupon two things happen :

(1) The thread or fibre becomes freely extensible ;

(2) ATP is hydrolysed and the thread or fibre contracts, doing work and developing tension.

It may be supposed that in the living muscle, the prosthetic groups of the myosin are normally inactivated or " masked " in some way ; since ATP is present, the actomyosin filaments are extensible. On excitation, the prosthetic groups are unmasked, or activated, ATP is hydrolysed and the muscle contracts. After a short time (the duration of the " active state "), the prosthetic groups become masked again, and since ATP is still present (only a small fraction need have been hydrolysed) the actomyosin regains its extensibility and the muscle relaxes. Since there is no change in the ATP concentration of a muscle, even after it has been made to contract repeatedly until it is fatigued to exhaustion (Table 16. 1), it must be supposed that there is an auxiliary reaction which synthesises ATP. After death, ATP is no longer synthesised, its concentration falls to zero and accordingly, the muscle becomes stiff and inexcitable. This *rigor mortis* can be reversed by adding ATP.

Creatine phosphate is also an " active " (" energy rich ") phosphate compound and its phosphate group can be transferred to ADP forming ATP, so :

$$\mathrm{CrP} + \mathrm{ADP} \rightleftharpoons \mathrm{Cr} + \mathrm{ATP}$$

This is the first of the reactions by which ATP is synthesised rapidly during and immediately after a contraction, and its concentration maintained. When the muscle is fatigued to exhaustion, the concentration of creatine phosphate falls very markedly (Table 16. 1) ; but in more ordinary circumstances this does not occur, since there are still further reactions by which creatine phosphate is re-synthesised.

Glycogen is the most important of the substances concerned in these reactions. In the presence of oxygen it is oxidised to carbon dioxide and water and, as described in Chapter 6 (p. 204), a large fraction of the successive steps in the whole reaction sequence are coupled with the synthesis of ATP. The creatine phosphate concentration will then be restored by a reversal of the reaction given in the previous paragraph, by which the temporary deficit of ATP was made good by a loss of creatine phosphate.

As pointed out in Chapter 6, the reaction sequences by which proteins, fats and carbohydrates are oxidised all converge on a common pathway ; and most of the steps in this pathway are associated with the synthesis of ATP. *Oxidative* recovery of a muscle, therefore, can be brought about by metabolites other than glycogen. Skeletal muscles, however (of both frogs and mammals), apparently use carbohydrates only, since the respiratory quotient of the excess metabolism (above the resting value) associated with muscular activity is 1·00. Cardiac muscle, on the other hand, seems to be less exacting, since the respiratory quotient is slightly, but definitely, less than 1 ; mammalian hearts (but not frog's hearts) can probably use fats ; frog's hearts can use amino-acids, since there is a steady excretion of urea and ammonia into the fluid perfusing an isolated heart.

The rate at which glycogen (or any other metabolite) can be completely oxidised is limited by the rate at which oxygen can be supplied by the blood stream ; full oxidative recovery takes a long time, even after a relatively short period of activity. There are, however, two ways in which the concentrations of creatine phosphate and ATP may be temporarily restored at rates which are not limited by the immediate rate of supply of oxygen.

One occurs only in those kinds of muscle which contain *myoglobin*. As pointed out in Chapter 3 (p. 80) myoglobin will take up oxygen at a partial pressure at which hæmoglobin has lost most of its oxygen, and will give it up again at a partial pressure adequate for the tissue oxidative enzymes. Myoglobin, therefore, will be able to supply these enzymes with oxygen during and immediately after each contraction of the muscle provided, of course, that it is able to recharge itself subsequently. That it actually does so can be seen from the changes in the absorption spectrum of the muscle. Appreciable loss of oxygen from the myoglobin occurs in 0·2 sec. after the beginning of a tetanus, and 60 per cent. saturation is reached in less than 1 sec. even when the blood supply is intact. The reserve of oxygen, therefore, is exhausted during the first second or so of the tetanus, after which myoglobin can play no part in the supply of oxygen. In intermittent short contractions of individual fibres at relatively long intervals, as occurs in tonic contractions, for example, the myoglobin may well be able to supply the heavy but intermittent demand for oxygen after each contraction. Cardiac muscle contains considerable quantities of myoglobin. It has been calculated that the amount present in a mammalian heart is just about sufficient to provide the oxygen required by a single maximal contraction—*i.e.* for about 0·5 second ; at moderate activity, the supply would last for about 7 seconds.

Lactic Acid Production. In the absence of a sufficient supply of oxygen to complete the breakdown of glycogen, the reaction sequence is broken at an intermediate stage, and lactic acid is formed (Chapter 6, p. 202). This reaction yields energy and can be coupled to the formation of ATP. The immediate sources of energy are thus restored but, as shown in Table 16. 1, lactate ions accumulate and the acidity of the

intracellular fluid tends to increase progressively. If the muscle is *in situ* in an animal, the acid is removed and buffered by the blood stream (as described in Chapter 3, p. 84) ; but if the muscle is isolated, the acidity eventually reaches a value which prevents the muscle from working (as shown in Fig. 16. 14), and it becomes fatigued to exhaustion.

When oxygen becomes available, frog's muscles not only oxidise the lactic acid in place of glycogen, but re-synthesise the glycogen lost. Mammalian cardiac muscle can use lactic acid in place of glycogen, but mammalian skeletal muscles can do so to a limited extent only, and most of the lactic acid produced passes into the blood and then to the liver where it is used to synthesise the liver glycogen. The muscle glycogen which is lost by oxidation or converted into lactic acid is replaced by synthesis from the glucose supplied in the blood ; this, in turn, is derived from the liver glycogen. This cyclic process is illustrated in the diagram in Chapter 6 (p. 196). It appears that the only way in which muscles can get energy anaerobically (apart from the use of the store of creatine phosphate) is by turning glycogen into lactic acid ; the consequent increase in acidity is an unfortunate necessity. There appears to be no way at all of getting energy anaerobically from proteins or fats.

When muscles are treated with iodoacetate, the breakdown of glycogen to lactic acid is inhibited. But in spite of this, skeletal muscles can still perform a great deal of mechanical work, even in anaerobic conditions, by making use of the stores of creatine phosphate. When these are exhausted, the muscles become inexcitable and go into a state of contracture, consistent with the exhaustion, also, of the store of ATP. If a frog's heart is treated with iodoacetate, it continues to beat indefinitely so long as oxygen is present. But in anaerobic conditions, it fails after 20 to 30 beats and is then found to have used up all its creatine phosphate ; the normal heart contains only about one-tenth as much creatine phosphate per unit weight as does a frog's skeletal muscle, It was these actions of iodoacetate that focused attention on the importance of creatine phosphate in the anaerobic recovery process.

Visceral (unstriated) muscles remain active for a considerable period in the absence of oxygen, spontaneous activity and tonus being lost more rapidly than the response to external stimulation. In partially, or completely, anaerobic conditions, they have been found to produce lactic acid from glycogen and, on the whole, they seem to behave in much the same way as skeletal muscles.

Mammalian cardiac muscle, on the other hand, is very sensitive to lack of oxygen and fails rapidly if the supply is inadequate for the recovery process. The " lactacid " mechanism is thus of little significance. Frog's hearts, however, survive well in anaerobic conditions, making use of energy released in the conversion of glycogen to lactic acid. It is essential that the lactic acid should be neutralised as rapidly as it is formed. If the pH of the perfusing fluid is kept at about 8·5, a frog's heart will go on beating in the absence of oxygen for 2 or 3 hours ; it stops eventually as a result of lack of carbohydrate for conversion into lactic acid, for if at this stage glucose is added to the perfusion fluid, there is an immediate and quite dramatic recovery.

"Oxygen Debt" in Man

If the oxygen consumption of a man is measured continuously by any of the methods described in Chapter 6 (p. 179), and he suddenly starts doing muscular work, it is found that the rate at which oxygen is consumed does not rise suddenly, but gradually, over a period of 1 or 2 minutes, up to the value which corresponds with the rate of working. Conversely, when he stops working, the rate of oxygen consumption falls gradually to the resting value. The total quantity of oxygen used in the recovery period is equal to the deficit (or "debt") which he incurred at the beginning of the exercise. External work, therefore, can be performed at the expense of energy stored in the body, and not only from that derived from a simultaneous consumption of oxygen. After the end of the exercise, these stores are replenished.

Additional evidence for the existence of these stores comes from the fact that *short* spurts of exercise can be performed at rates which greatly exceed the limiting rate of *long-continued* exercise, which is set by the rate of oxygen intake and the output of the heart. In such conditions, part of the energy needed for the muscular work—and possibly quite a large part—must have been stored ; the man incurs an "oxygen debt" and the intake of oxygen is postponed until the period of recovery. A. V. Hill's "ordinary middle-aged University Professor," for example, could incur an oxygen debt of 5·5 litres in 13 seconds by running 100 yards at top speed ; that is, he consumed 5·5 litres of oxygen over and above his resting oxygen consumption during the period of recovery from the sprint. His oxygen requirement during the period of exercise was thus at the rate of 25 litres a minute, whereas his maximum rate of oxygen intake was found to be only about 4 litres a minute. Consequently he exerted himself at least six times as strenuously, for a short time, as he could have done if his supply of energy had been limited to that obtained from current oxidation in the body. The power exerted by an athlete in a 100 yards sprint may be ten times as great as that of a long-distance runner whose maintained power output is limited by the rate at which oxygen can be absorbed and distributed by the blood ; and he may have to use as much as 15 to 20 litres of oxygen in repaying the debt incurred while he was running.

The evidence discussed in the previous section, derived chiefly from studies on isolated muscles, suggests that the production of lactic acid from glycogen, and to a smaller extent the creatine phosphate reserve, constitute a store of energy which can be released without the use of oxygen. Does this apply also to an intact man, and is the repayment of an oxygen debt an indication that this store is being replaced ?

That an oxygen debt is associated in man with the production of lactic acid is shown by the observations plotted in Fig. 16. 15. If the oxygen debt is greater than about 4 litres, the lactic acid concentration in the blood rises in direct proportion to the oxygen debt incurred ; if the oxygen debt is less than about 3·5 litres, the lactic acid concentration

does not rise. This suggests that there are creatine phosphate reserves which are used up first, and when they are depleted, lactic acid is produced. The amount of creatine phosphate in a resting skeletal muscle amounts to slightly more than 5 g. per kg., and each gram can liberate on hydrolysis about 0·04 kilocalories. The average man has about 30 kg. of muscles, so that he has, altogether, about 150 g. of creatine phosphate which could yield a total of about 6 kcal. of energy without needing oxygen. As indicated in Fig. 16. 15, following the most severe exercise, the lactic acid concentration in the subject's blood increased

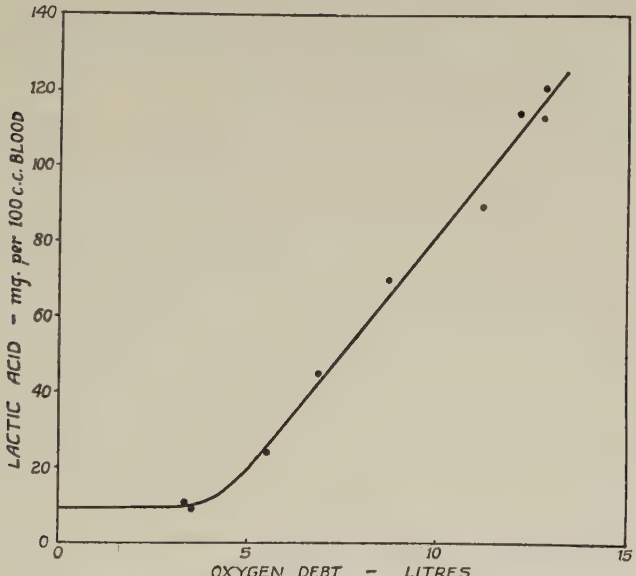


FIG. 16. 15. Showing that **lactic acid** accumulates in the blood during exercise when the **oxygen debt** increases above the limit set, probably, by the exhaustion of the creatine phosphate stores.

The subject, a young athlete, took ten minutes' steady exercise of different degrees of severity, and the extra oxygen absorbed by him during recovery (that is, used to pay back the oxygen debt incurred) was compared with the lactic acid concentration in his blood immediately after the exercise. (After Margaria, Edwards and Dill.)

by 110 mg. per 100 ml. blood. If we assume that the lactic acid was distributed more or less uniformly throughout all the water in his body—*i.e.* in about 45 litres (the subject weighed 61·2 kg.)—the total quantity present must have been about 50 g. The energy liberated when 1 g. of lactic acid is formed from glycogen is 0·33 kcal., so that the total amount of energy liberated by the "lactacid" mechanism was about 16 kcal.

In order to replace these stores, energy must be supplied from another source. The amount necessary is at least as great as the amount released when the stores break down, and may be expected to be appreciably greater: no practical engine, living or not living, can operate

with an efficiency of 100 per cent. or it would be a " perpetual motion machine." As we have seen in Chapter 6, each litre of oxygen consumed liberates about 5 kcal. of energy whatever the foodstuff burnt. The consumption of nearly 4 litres of oxygen in repaying the oxygen debt incurred without production of lactic acid, and of an additional 9 litres in repaying the maximum oxygen debt incurred, would yield about 18 kcal. and 45 kcal., respectively. The amounts of energy available are thus about three times the minimum amounts necessary to restore the estimated creatine phosphate reserve, and to re-synthesise the glycogen from which the lactic acid was derived. There is no reason to suppose that the stores of energy used by a man when he contracts an oxygen debt differ essentially from those used by an isolated muscle of a frog during the anaerobic recovery process.

CHAPTER 17

RECEPTORS AND SENSATIONS

MECHANICAL STRAIN, TEMPERATURE, TASTE AND SMELL

No animal can act in a purposive way or control its own bodily functions unless it is supplied with enough information about the outside world and the state of its own body ; it is the function of the receptors to supply this. Much of the information received by the central nervous system from receptors sets up reflex responses ; for example, receptors in the muscles of the limbs are continuously signalling the lengths of these muscles and in the absence of other activity they are kept, by reflex action, at a constant length. In ourselves, and by inference in other persons and in the higher animals generally, some of this information also reaches consciousness and underlies sensation.

Our own experience of receptor mechanisms is dominated, as was the early experimental work on the subject, by sensation. It is therefore important to understand the relation between a sensation and the activity of the receptor organs. This relationship is probably best illustrated by describing briefly the events that are going on when we are conscious of a sensation. Let us consider a visual example : electro-magnetic waves of widely varying wave-lengths are reflected from various parts of the object seen, different wave-lengths being reflected with greater or less intensity from the different parts ; those wave-lengths within a certain narrow band, which we call light, are able to excite receptors in the retina and these set up nerve impulses ; the activity in any one nerve fibre is combined with, and modified by, the activity of other nerve fibres at the junctional regions between them ; several such junctional regions, at each of which changes in the message will occur, are to be found before the information reaches the cerebral cortex where, it is generally presumed, the entirely unexplained processes involved in consciousness and sensation take place.

This illustrates certain things about a sensation. First, it is and can only be a subjective response. Secondly, a sensation cannot include more information about an object or event than has been transmitted by the physical events outside the body and by the activity of the receptors and the nervous system. Thirdly, sensation may be a product, not only of the information from one particular kind of receptor, but also of that from other kinds of receptor of the same general type but differing in detail. Fourthly, the intensity and quality of a sensation depend on the nature and intensity of other sensations, on the general activity of the central nervous system at the time, and on its antecedent activity associated with previous experience. One can diminish a sensation of pain by applying a counter-irritant ; one can fail to hear

quite loud sounds if one is deep in thought ; and one can fail to observe, by touch or sight, a familiar object in its accustomed place. It is this discriminating synthesis by the central nervous system of the information transmitted to it by a variety of receptors which forms the basis of the wide range and subtlety of sensation of which the human mind is capable. Sensations depend as much on the central nervous system as on the receptors, and great care must be taken to scrutinise the interpretation when sensation is used as an index of receptor activity.

The most striking thing about our sensations is their clarity and detail, but it must remembered that there are real limitations ; for example, " sounds " with a frequency greater than 20,000 c/s are inaudible, and the effect of doubling the intensity of a sound wave is not to increase its loudness by a factor of 2, but by a variable factor, which is quite different for high and low notes. It is the purpose of this chapter to consider how the receptors are able to transmit so much information to the central nervous system and to describe also some of their limitations.

The Behaviour of Receptor Units

The first step in considering the problem of the transmission of information to the central nervous system is to consider the behaviour of the individual unit ; the word unit being used to describe a single afferent (sensory) axon and the one or many receptors with which it is directly connected. It seems probable that all types of receptor, including the receptors in the complex organs of the so-called " Special Senses," behave in fundamentally the same way ; however, most experiments on single units have been done in places other than the organs of " special sense."

The Specificity of Receptors. Single afferent nerve fibres can be dissected out from nerve trunks and separated from other fibres which have different connections. When this is done, it is usually found that activity appears in response to excitation of the receptors from which the fibre is derived ; and usually, it appears much more readily when one particular type of energy is used, *e.g.* mechanical, thermal, chemical or light, than when other types are used. Thus the receptors of the retina normally respond only to light, and those associated with the senses of taste and smell respond only to chemical stimulation. In general, therefore, each kind of receptor is specially sensitive to one form of energy : receptors are *specific* in that each unit exhibits a particular pattern of sensitivity (Muller's Law). One class of receptor which does not show big differences in its sensitivity to different forms of energy but shows a distinct sensitivity pattern of great importance, is that associated with the sensation of pain. These units respond to high intensities of stimulation by all types of energy—that is, levels which are nearly or actually damaging. When receptors, which are specially sensitive to one form of energy, are excited by a high intensity of some other form, the sensation aroused is that expected of the normal stimulus, demonstrating that the sensation is dependent on the

connections of the unit and not on the nature of the stimulus that excited it. A good example of a sensation being aroused by an abnormal stimulus is to be found in the flash of light that is seen if the eyeball is struck, or if an electric current is passed through it.

The Response of a Single Unit. The response that can be recorded from a single afferent fibre when the area it innervates is stimulated, consists of from one to many impulses. These impulses, like all nerve impulses, are all-or-nothing; it is clear, therefore, that in any one fibre information can only be carried by variations in the duration of the intervals between successive impulses and in the total duration of the impulse discharge. The simplest type of response is found in units that signal a steady state; these include temperature receptors, some

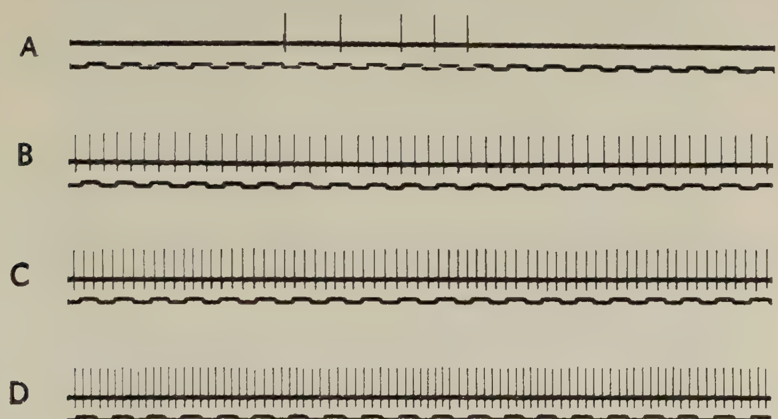


FIG. 17. 1. The relation of frequency of discharge to stimulus strength. Records of impulses from a single fibre of the incisor nerve of an anæsthetised rabbit while different weights were hung from the tooth. A, 5 g. (threshold). B, 10 g. C, 20 g. D, 50 g. All records one minute after the weight was applied. Note occasional irregular impulses in A. Below each record are time marks in 1/10th second.

All the action potentials are seen to be of the same size, irrespective of their frequency and thus of the strength of stimulation, illustrating the all-or-nothing law. (From a record kindly prepared by Mr. A. R. Ness.)

mechanical receptors, some receptors in the retina and the chemica receptors associated with smell. Such a receptor, when in a steady state, sets up a train of impulses which lasts as long as the stimulus; the frequency of the impulses depending on the strength of the stimulus (Fig. 17. 1; compare also Fig. 4. 20, p. 127). These frequencies are repeatable so that, for example, whenever a limb is in a given position, a given receptor in the joint capsule will discharge at a particular frequency. This is important, since the nervous system can only act consistently if the information it receives is consistently related to the stimulus.

The response is not quite as simple as that just described if the stimulus is applied abruptly and then left constant; for example, if

a muscle stretch receptor (muscle spindle) is suddenly stretched to a new length, the impulse discharge starts at a high frequency, which declines at first rapidly and then more slowly until the steady frequency associated with the particular degree of stretch is reached. This decline of frequency after the onset of a steady stimulus is called *adaptation*. The frequency at the beginning of such a discharge depends not only on the final size of the stimulus, in this instance the amount of stretch, but also the rate at which the final level is reached, in this example, the velocity of the stretch. Discharges from temperature receptor units which behave in this way are shown in Fig. 17. 5 (p. 515 below).

A large number of receptors do not respond to steady stimuli at all. They respond to a suddenly imposed steady stimulus by a discharge at a frequency which is initially high but which falls rapidly to zero (Fig. 17. 2) ; and often give another burst of impulses when the stimulus

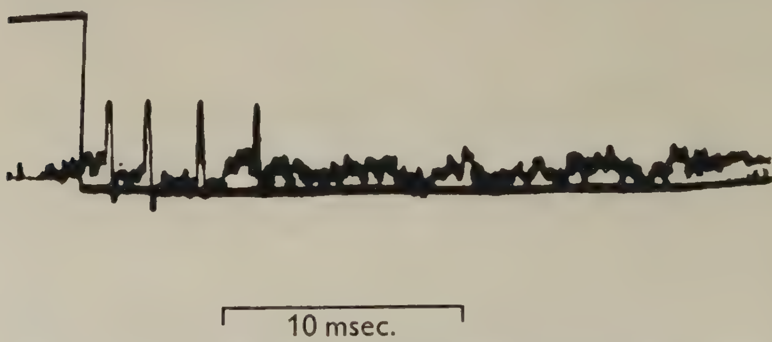


FIG. 17. 2. Rapid adaptation of a receptor discharge. Impulses recorded from a mesenteric nerve, while a single Pacinian corpuscle was stimulated by the movement of a piezo-electric crystal. The top trace (at left) indicates the amplitude and time course of the stimulus. (Gray and Matthews, *J. Physiol.*, 1951.)

ends. As an extreme case, the large capsulated end organs, known as Pacinian corpuscles, often respond to a steady pressure with a single impulse at the moment the pressure is applied and another when it is released. Such receptors which respond only to a change of state are called " rapidly adapting," or *phasic receptors* : they are to be found both among those responding to mechanical stimuli and those responding to light stimuli (Fig. 18. 18, p. 559).

A little thought about one's own sensations will immediately make clear the importance of signalling changes of state ; one responds vigorously to sudden changes of sound, while being oblivious of a steady noise ; and if one wants to feel an object carefully one keeps one's hand moving and allows the rapidly adapting receptors to respond as the hand passes over the small irregularities of the surface.

Receptive Fields. A single nerve fibre may come from a single receptor or from many ; if there are many, they may be spread over a wide area. The receptive fields of primary receptor fibres may there-

fore vary from a single point up to quite considerable areas ; in the cat there are units sensitive to mechanical deformation which have receptive fields up to 5 cm. by 9 cm. in size. Receptive fields usually overlap.

Receptor Mechanisms

It is probably true to say that all receptors when excited set up changes in electrical potential, which, unlike the nervous impulses, are not all-or-nothing, but vary in size and duration with variations in magnitude, velocity and duration of the stimulus (Fig. 17. 3). These

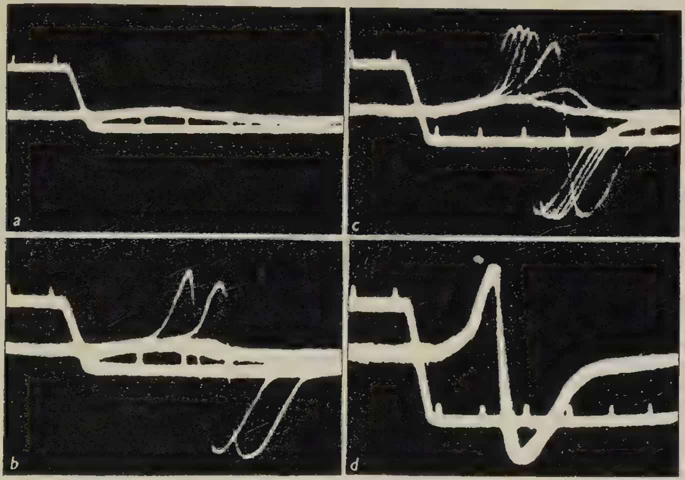


FIG. 17. 3. Receptor potentials and impulses in a Pacinian corpuscle. Each picture is of twenty superimposed traces ; the top beam (at the left) signals the amplitude and time course of the mechanical stimulus that was applied to the corpuscle and also shows 1 msec. time intervals. The stimulus strength was increased between each record. The receptor potential can be seen to increase in size with the stimulus and the impulses increase in number. In (*b*) and (*c*) 2 and 5, respectively, of the stimuli gave rise to impulses, which " took off " at different values of the receptor potential, owing to slight variations of the threshold. In (*d*) every stimulus results in an impulse. (Gray and Sato, *J. Physiol.*, 1953.)

Receptor Potentials, which are localised in the receptor or its neighbourhood, are the immediate cause of the impulse discharge. The ways in which the various types of stimulus set up these potentials are not yet known with any certainty. However, whatever the mechanisms involved in setting up these potentials, they must be highly sensitive ; a rod cell in the retina can be activated by a single light-quantum, *i.e.* the smallest possible quantity of light ; movements of the basilar membrane of molecular or even atomic dimensions are adequate to set up impulses in response to sound ; the threshold movement on the outside of a Pacinian corpuscle (not on the sensitive element) is probably about 0·2 μ—below the limit of resolution by the best microscope.

In view of this high sensitivity, it is not surprising to find that the energy received by a receptor, in the form of light or mechanical work, for example, is too small to account for the electrical energy which is released by the receptor, and which can be recorded as the receptor potential. The external event must, therefore, release energy from an internal store. It appears that receptors store energy in the form of concentration gradients of ions across membranes. The stimulus causes a change in the permeability of this membrane and allows the ions to move down their electrochemical gradients. In general, this process is similar to the release of electrical energy in the form of end-plate or synaptic potentials by a chemical transmitter.

The Transmission of Receptor Information

How can receptor units when working in large numbers provide enough information for complex sensations? Information on the strength and duration of a stimulus may be conveyed by the time course of the discharge in each fibre, but information about stimulus strength is also signalled by the number of units activated. The time course of the discharge can, however, convey nothing about the site of the stimulus or its nature; such information reaches the central nervous system through the organisation of particular fibres, each of which has its own particular properties. There are thus two general ways in which information is signalled: in time, and by distinguishing between different units; or one can say in time and space. These two general ways are correlated with the two types of summation found in the central nervous system, temporal summation between successive impulses and spatial summation between impulses in neighbouring fibres.

It is now possible to summarise the variables available for transmitting information :—

Factors depending on time.
- (*a*) The intervals between impulses.
- (*b*) The duration of the discharge.

Factors depending on the specific properties of the receptor units.
- (*a*) The specificity of the unit to a particular type or types of stimulus and its rate of adaptation, which decides whether it is sensitive to a steady state or a change of state.
- (*b*) The size of the receptive field and its position in the body.
- (*c*) The sensitivity of each unit and, arising from this, the number of units active for a given stimulus.

Thus the time intervals signal the strength of a stimulus, but information as to whether this is intensity of illumination, rate of change of illumination, amplitude of displacement, or velocity of movement, depends on what type of unit is responding. It will be noticed that stimulus strength may also be signalled by the number of fibres active at a time.

To illustrate these points, let us consider the eye: some receptor

units, particularly at the periphery of the retina, are specially sensitive to any change (in either direction) of light intensity, while others signal the steady intensity : the pattern can be transmitted because each receptor has a particular place in the surface on which the image is focused ; the intensity or rate of change of intensity is signalled by the frequency of the discharge in each unit and the number of units responding ; the colour can be seen because of the specificity of certain receptors to certain wave-lengths.

SENSATIONS ASSOCIATED WITH MECHANICAL STRAIN

There are many kinds of receptor which are activated and discharge nerve impulses when they are distorted or deformed. The deformation may be produced by contact with objects in the outside world, by changes in the relative position of different parts of the body as a result of muscular contraction, or of movements and changes in orientation with respect to gravity. The stimuli to which they respond are thus essentially mechanical, so that it is convenient to group them together. Receptor units which can respond to mechanical stimulation are found in the skins or other integuments, in the muscles, and in the joints or other articulations, of species throughout the animal kingdom. There are many varieties of such units, and many of them are associated with special structures which make them particularly sensitive to certain kinds of mechanical deformation. The Pacinian corpuscle is one such structure, of a relatively simple kind, and others will be mentioned in the following paragraphs. In the ear, the associated structures are so elaborate that the receptor cells—which are not in themselves essentially different from other mechanically sensitive receptors—respond only to the changes in air pressure which constitute sound ; and the whole arrangement makes up an organ of " Special Sense."

Touch and Pressure

The sensations experienced by human beings when they come into contact with objects outside them are described as " touch " or " pressure." Such terms have great value in any investigation in which a human sensation is used as a criterion, but are of little value when discussing the physiological processes which underlie the sensation. The information on which these sensations are based is received by receptors which are sensitive to mechanical stimuli, and which lie in the skin. The receptor units, even in one individual, vary in all the ways that have been mentioned in the last section : they vary in their sensitivity to the direction of the force ; they vary in the time course of their responses and, associated with this, they vary in their responses to different rates of change of the applied force ; they vary in the size of their receptive fields ; and even when these factors are the same, they may still vary in their sensitivity to a force of given magnitude. The specialisation of structure which can be seen under the light microscope, and is found among receptors in the higher mammals,

may be associated with particular functional properties. It is certain, however, that great variation in the properties of receptor units can be found in species that do not show any gross morphological specialisation of receptors ; and even in those that do—man, for example—there are areas which contain no receptors specialised in this way but which are, none the less, associated with a full range of sensations. In any event this is not a matter of great importance in this context. What does matter is that there are, functionally, many types of receptor unit each with its own quantitatively characteristic properties.

This variety in the properties of receptors is probably essential to the full development of all the qualities of any kind of sensation ; but it is none the less often possible to associate the main characteristics of a sensation with a type of receptor having certain properties. An analogy is that of a musical instrument which plays a certain note ; we refer to this note by a letter which describes the frequency of the fundamental, but the quality that is characteristic of the instrument is determined by the harmonic content of the note. If we draw a piece of cotton wool lightly over hairy skin (a test used in clinical medicine) we are conscious of a particular type of the sensation of touch. We can arouse the sensation when only the hairs are touched, and the sensitivity of the area is much reduced if it is shaved. There are receptors associated with the hair follicles, and comparable receptors in experimental animals have been shown to have the high sensitivity and rapid adaptation which are also associated with the sensation : it seems reasonable, then, to associate the basic part of the sensation with these receptors. It must be pointed out, however, that the receptors themselves need not adapt rapidly, even though the associated sensation adapts rapidly ; there are reasons for believing that certain sensations adapt as a result of central processes. A sensation of touch can also be aroused on non-hairy skin, such as the palms of the hands. " Touch " implies that the sensitivity must be high and that the receptors are so placed that they can provide precise information about localisation of the stimulus ; such receptors are likely to be superficial. " Pressure," on the other hand, implies a rather stronger stimulus over a fairly wide area, and since the sensation does not adapt, neither can the receptors.

Being sensations, touch and pressure are the products of the activity of the central nervous system as well as of that of the various kinds of receptor just described. Some of the attributes of such sensations can even be lost by lesions of the nervous system without affecting others ; with certain lesions, for example, localisation of the sensation may be lost, while sensitivity is retained.

Proprioception

Information about the position of the body and of its parts is provided by mechanically sensitive receptors of a variety of types which are stretched or deformed in various specific conditions. *Muscle spindles* provide information about the lengths of muscles ; *tendon*

receptors provide information about the tension set up by the muscles; receptors in *joints* indicate the angle or angles between the bones. Most of these receptors are able to signal a steady state, although they are all more active during a change. Information from these sources, together with that from receptors in the labyrinths (see below) and from the eyes, is used by the nervous system to control posture and movement: this information can also form the basis of a conscious appreciation of the orientation and movement of the body in space and the position of the limbs with respect to the rest of the body.

Lastly, the *baroreceptors* in the aorta and carotid sinus, sensitive to the arterial blood pressure (Chapter 2, pp. 42–47), and the stretch receptors in the *lungs* (Chapter 4, pp. 126–128) are all receptors sensitive to mechanical strain. The information provided by these is important for initiating cardio-vascular and respiratory reflexes, but rarely, if ever, forms the basis of conscious sensation.

The Labyrinths

Each labyrinth lies in the petrous bone in close anatomical relation with the inner ear, with which, however, it has no physiological connections (Fig. 19. 3, p. 570). It consists of three specialised structures, each containing mechanically sensitive receptors. (1) The *utricle* encloses a small calcareous body (the *otolith*) lying against the projecting filaments of the receptors (hair cells). The receptors are activated by the weight of the otolith and so provide information as to the orientation of the head with respect to gravity; they are also activated by linear accelerations. (2) The *saccule*, which also contains an otolith, is perhaps also concerned with signalling the position of the head in space. In the frog, the saccule is concerned only with the detection of low frequency vibration. In the cat, there are apparently two separate systems which signal tilts from side to side, and fore and aft, respectively, but they have not yet been identified anatomically. (3) The *semi-circular canals*, lying in three planes at right angles to each other, and filled with endolymph. Each canal contains a *cupola*, which consists of hair cells whose bases are embedded in a gelatinous matrix in which they are in contact with a number of nerve endings. When the head is rotated, the endolymph in the semi-circular canals, owing to its inertia, lags behind; the cupola is dragged through the endolymph and thus displaced and distorted.

Electrical studies have shown that the nerve endings in the cupola discharge spontaneously when the head is at rest. When the cupola is displaced in one direction the frequency of discharge is increased, and when it is displaced in the opposite direction, the frequency is decreased (Fig. 17. 4). As a result, short-lasting rotations are accurately signalled. In steady rotations, the discharge of impulses behaves as though the cupola were displaced during the period of acceleration, and then drifted slowly back to its resting position during the period of steady rotation. Suddenly stopping the rotation of the head will displace the cupola in the opposite direction, and although movements

of the endolymph die down in about three seconds, the cupola seems to take about twenty-five to thirty seconds to return to its resting position. During this time the subject will experience a sensation of rotation in the opposite direction.

Reflexes based on information provided by the receptors in the labyrinths are important, not only in the maintenance of posture (as discussed in Chapter 14, p. 416), but also in controlling the position of the eyes, as will be discussed in the next chapter.

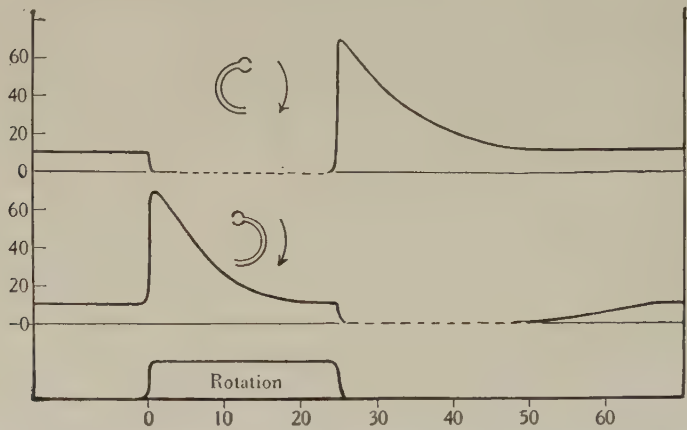


FIG. 17. 4. Frequency of **Action Potentials** from the horizontal **Semicircular Canal** of a cat, when acceleration and deceleration are separated by an interval of steady rotation.

There is a steady discharge during rest which is suppressed by rotation in one direction, and augmented by rotation in the opposite direction. Cessation of the rotation leads to an after discharge, or a silent period according to the direction of the rotation. (Adrian.)

TEMPERATURE

Much of what has been said in the previous sections about the relationship between receptor units and sensation is also applicable to that between the receptor units affected by temperature and the sensation of temperature. Temperature receptors are found in all parts of the skin and in some mucous membranes. Certain single units investigated in detail have been found to set up a steady discharge whose frequency depends solely on the temperature of the receptor; for a given receptor there is a temperature at which the frequency of the discharge is a maximum, and the frequency declines at temperatures on both sides of this maximum. There are two main groups of receptor, those which have a maximum discharge around a temperature of 30° C., and those which have a maximum around 40° C. A particular combination of frequency of discharge in each group of units is thus uniquely related to a particular value of temperature. Information from a single group of receptors having a temperature–frequency character-

istic of the type mentioned would always be ambiguous, since it gives the same frequency of discharge at two separate temperatures. When a man is in a " comfortably warm " state, the temperature of his skin ordinarily lies between 25° and 35° C., that of the temperature receptors being a few degrees higher : in these conditions, a rise of temperature causes one group of receptors to discharge more rapidly, and the other group to discharge less rapidly, while a fall in temperature has the converse effect. Both groups of receptor respond with a burst of impulses, over and above the steady response, if the temperature is suddenly changed ; the " 30° " receptors respond in this way to a sudden fall but not normally to a sudden rise of temperature, and the " 40° "

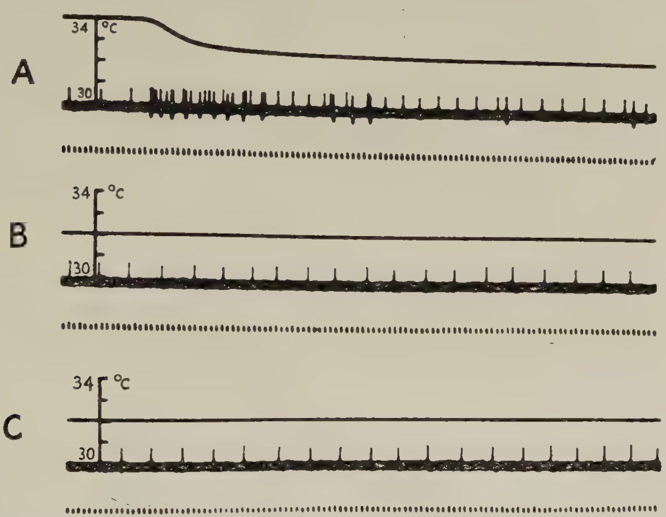


FIG. 17. 5. The impulse discharge, in two nerve fibres from **temperature receptors** in a cat's tongue, showing **adaptation.** The upper line signals the temperature, which falls at the beginning from 34° C. to 32° C. At the bottom are time marks in 1/50th second. The discharge in one fibre (diphasic potentials) adapts to zero, while the other (monophasic potentials) adapts to a steady frequency. The records were taken at the following times : A, zero ; B, 1 minute ; C, 15 minutes. (Hensel and Zotterman, *Acta physiol. Scand.*, 1951.)

receptors respond to a sudden rise of temperature. These higher frequencies resulting from change of temperature decline rapidly and within seconds, or at most one or two minutes, have reached their steady level (Fig. 17. 5).

Certain other points of interest arise if we consider the sensations aroused from human skin as a whole instead of considering the responses of single units from experimental animals. If an area of skin is tested for these sensations using metal rods with small tips (say, diameter 1·5 mm.), which feel hot or cold, it is found that there are certain areas especially sensitive to either the sensation of warmth or that of cold. In these areas the sensation is easily obtained in the centre, but less easily in the immediate surroundings. Provided the testing rod is left

long enough in contact with the skin, temperature differences can be felt on most parts of the skin owing to the conduction of heat to or from the sensitive spots. The areas responsible for the sensation of warmth do not coincide with those responsible for the sensation of cold ; neither are they equal in number. Certain " mucous membranes " are relatively insensitive to stimulation by raising the temperature, so that it is possible, for instance, to drink liquids which are hotter than the skin can bear, and the same also applies to medicinal douches.

If one hand is placed in hot water and the other in cold, and if they are allowed to remain until the sensations of cold and warmth have subsided and then both hands are placed in tepid water, the hand that was in the hot water will now feel cold and the one that was in the cold feel hot. This effect is probably not due simply to an initial adaptation of the receptors themselves. Like some other peculiarities of the sensation of temperature, it may be due partly to the cutaneous vasomotor changes that accompany sudden changes in the temperature of the surroundings. The temperature of the deeper layers of the skin is rarely the same as that of the superficial layers, and the magnitude, and even sign, of the difference depends largely upon the rate of blood flow in the various parts of the cutaneous circulation. The temperature receptors in the two hands may thus be at different temperatures, even though both hands are in water of the same temperature. But it is possible, also, that there is some central adaptation ; the sensation at any moment will not be determined solely by the signals sent in by the receptors at that moment.

PAIN

The sensation of pain is primarily based on information signalled by specific nerve fibres coming from receptors which respond to strong stimulation, whether mechanical, thermal or chemical. Excessive stimulation of other types of receptor cannot be presumed to produce pain, as it is possible to obtain very high impulse frequencies in large numbers of receptor units with stimuli that are not painful. This is not to say that excessive stimulation of more sensitive units cannot give rise to discomfort, or that these more sensitive units are not concerned in the quality of the sensation. Again, the sensation is the final output of the nervous system based on the available information : the basic requirement of a sensation of pain is a high intensity of stimulation, and the most important means available to the nervous system of signalling intensity over wide ranges is in the differences in sensitivity of the receptor units. Thus the high threshold units are the fundamental ones for this sensation.

Two types of pain, fast and slow, are usually described. This distinction was first made from results of subjective experiments, but analysis of the responses in single primary receptor fibres has shown that receptor units with a high threshold for all types of energy fall into two groups with different velocities of conduction of the nerve impulses. The faster group have a conduction velocity comparable to that of the slower

low threshold mechanically sensitive units concerned in the sensations of touch and pressure ; while the slower group are among the slowest of all sensory fibres.

A special word must be said about visceral pain. Most of the information from the visceral receptors never reaches consciousness, but on occasion a sensation of pain may be aroused ; when it is, the pain may be referred by the nervous system to the skin belonging to the same segment as the visceral nerve responsible. This is known as "referred pain." For example, gall bladder pains are often localised in the right shoulder ; kidney pains in the groin. One finds that the receptor nerves to joints, to the muscles working those joints, and the covering skin are all supplied from the same spinal cord segment. In joint injuries it is usual to find impaired movement due to fixation by the muscles, and also pain in the overlying skin. Visceral sensations are usually painful and are of the same type whatever their origin. Two very common causes of visceral pain are (1) prolonged contraction of plain muscle such as occurs in a ureter when partly obstructed by a stone, and (2) the stretching of organs such as the mesentery.

TASTE AND SMELL

There are many different kinds of receptor which are sensitive to the presence of various chemical substances in the fluid around them. Many of these, such as the " chemoreceptors " in the brain and carotid body, concerned in the regulation of breathing (Chapter 4, p. 121), and the " osmoreceptors " concerned in the control of the concentration of the urine (Chapter 9, p. 273), are of great importance in the life of man, as well as of that of all animals ; but their activities do not give rise to any definite kind of sensation. Certain specific chemical substances, however, when present in or near the appropriate kinds of receptor, may arouse the sensations of taste and smell. In this respect, the two sensations are superficially related ; but when examined in more detail, it is seen that there are very considerable physiological differences between them.

Taste

The mucous membrane of the epiglottis and soft palate, and of the tip, sides and root of the tongue, contain special receptor organs known as taste-buds. In these are the receptors for the sense of taste. All tastes can be divided into four (or perhaps six) groups : the sour, the salt, the bitter and the sweet, to which are sometimes added the metallic and the alkaline. It is significant that although many substances give rise to mixed sensations of taste, it is nearly always possible to distinguish the components, and it is impossible to create an entirely new taste by combining any or all of the " pure " tastes. The sense of smell, as we shall see, differs very markedly in this respect. It must be remembered that in most cases the actual flavour of any substance present in the mouth depends upon its smell almost as much as on its taste *sensu strictu*.

Sourness is a sensation aroused by all solutions containing hydrogen ions in sufficient concentration. For the mineral acids, such as HCl, the threshold concentration is at a *p*H of about 4. The organic acids, such as acetic, and also carbonic acid, appear more sour than would be expected from their hydrogen ion concentration, probably owing to the greater ease with which they penetrate through cell membranes.

Saltness is a sensation aroused by the salts of the strong acids, particularly the monobasic acids. The least concentration of NaCl which can be tasted is about 0·02 M (0·12 per cent.).

Bitterness is a sensation aroused by many substances with a very wide range of chemical composition, but above all of the alkaloids such as strychnine and quinine. Salts of magnesium, calcium and ammonium have a bitter taste, and so have ether and most glucosides. The threshold concentration for strychnine is about 0·00006 per cent.

Sweetness is a sensation aroused by the sugars, and also by a number of other completely unrelated compounds, *e.g.* beryllium salts, lead acetate, chloroform, many amino-acids and saccharin. The least concentration of cane sugar which can be tasted is about 0·5 per cent., whereas that of saccharin is only about 0·001 per cent.

These four sensations are not equally easy to arouse in all parts of the tongue, some being most easily aroused in some parts, and others in other parts. There are even substances such as magnesium sulphate and dulcamarin (the glucoside from bitter-sweet) that give rise to different sensations when applied in different places. The receptors which supply the information on which these sensations are based must vary in their sensitivities to the different classes of substance ; but records from single receptor units in various species of experimental animal indicate that there is not a distinct class of receptor for each of the four sensations. Thus in the cat, there are receptors responding predominantly to acids ; others responding to acids and to salts ; still others responding to acids and to substances like quinine ; and some responding to sucrose. We do not know that the receptors in man have the same sensitivities as those in the cat, but such a pattern of sensitivities contains all the variability required to transmit the necessary information ; there is a unique combination that responds to each class of substance. This is another example of a principle often found in the receptor nervous system, already referred to in connection with the sensation of temperature.

Smell

The organs for smell are situated in the upper parts of the nasal cavity. Odorous substances in the inspired air dissolve in the mucus covering the sensitive cells, diffuse into the hairs which protrude from the cells into the mucus layer, and so excite the receptors.

In contrast with the sense of taste, it is quite impossible to classify the various types of smell into definite components ; each substance has its own distinctive smell. There are certain general resemblances, however, and it has been suggested that odorous substances can be

grouped into the spicy, the flowery, the fruity, the resinous or balsamic, the burnt and the foul. Unlike taste, again, the combination of two or more smells may produce a completely new smell, which cannot be analysed into its components. One smell, again, can mask, or neutralise another (the action of perfumes in this connection is well known), and this can take place even if the two odorous substances are applied to different nostrils.

One of the peculiarities of the sense of smell is its rapid "fatigue"; air which initially has a powerful smell may seem quite odourless within a few minutes. Recovery is equally rapid. This "fatigue" only applies to the particular substance exciting it, and another substance, even though it has a very similar smell, may be perceived normally. Different substances "fatigue" the sensory apparatus at different rates, but for any given substance, the rate of "fatigue" increases with the intensity of the smell. Some people are completely deficient in the sense of smell and many are incapable of smelling certain substances which have a strong odour to others (hydrocyanic acid is a typical instance). This deficiency may be congenital or acquired.

Action potentials in the olfactory bulb and the olfactory area of the brain of experimental animals have been recorded by Adrian. These potentials indicate the activity of cells on which information from thousands of primary units has converged; this activity, therefore, represents a stage in the analysis of the information. Nevertheless, it seems clear that different kinds of odorous substance may affect preferentially different groups of receptor. The number of such different groups appears to be very large, and each group may respond to a number of different substances; there is no indication of the existence of any analysis in terms of a small number of "standard" components. It was found, however, that on the whole water soluble substances, such as acetone or amyl acetate, excite preferentially receptors in the anterior part of the olfactory epithelium (in the rabbit), while oily (lipoid soluble) substances, such as cedar wood oil or benzene, excite receptors in the posterior part. The discharge set up by oily substances, also, had a longer latency, a less abrupt onset, and longer duration, than those set up by water soluble substances. The discrimination of smells may thus depend both on the spatial distribution of the activated receptors, and on the temporal characteristics of the response. The rapid "fatigue" to the smell of any one substance appears to arise in the central nervous system; the receptors themselves appear to be able to respond indefinitely.

CHAPTER 18

VISION

VISION in a man, as in other complex and highly developed animals, involves the detection of the light that falls upon him, reflected chiefly from objects in the surrounding world ; but it consists more particularly in the detection of the patterns of light, shade and colour formed by these objects, from which he forms a visual perception of the external

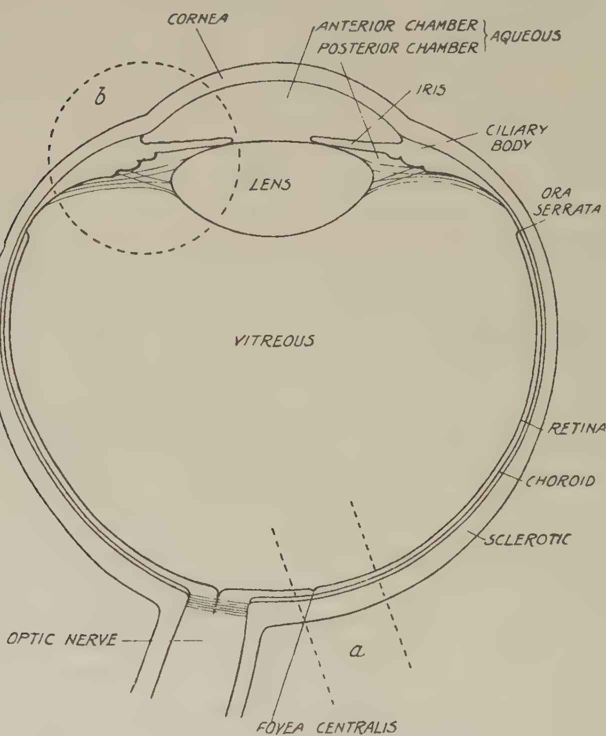


FIG. 18. 1. Diagrammatic Horizontal Section of the **Human Eye** (Parsons.)

world. The eye, accordingly, consists of an optical system by which an image of the external object is formed on a large number of closely packed light-sensitive receptor cells. These initiate patterns of impulses in appropriate fibres of the optic nerve, and consequently a spatial and temporal pattern of impulses in the visual cortex. The sensation of vision is presumably derived from this pattern.

The physiology of vision cannot be properly studied without some knowledge of the structure and properties of the eye. The outer coat

of the eyeball is called the *sclera* (the white of the eye), and this becomes continuous in front with the more highly curved *cornea* which is transparent to allow the entry of light into the eye. The shape of the combined cornea and sclera resembles that of a cricket ball with a watch-glass placed on it (Fig. 18. 1).

Acuteness of vision depends on the optical perfection of the anterior surface of the cornea. This structure is well protected by the eyelids and by the secretion of tear fluids. The lids themselves are composed of plates of compressed fibrous tissue, and they are involuntarily closed when a foreign body is seen to be approaching the eye or when the cornea is touched. Many kinds of animal have a third, semi-transparent, eyelid (the nictitating membrane) moving horizontally beneath the other two.

Tear fluid is secreted by the lachrymal gland ; it irrigates the conjunctival surfaces of the lids and cornea. In addition to its mechanical action of washing dust and other foreign bodies out of the eye, it has a bactericidal action. Normally the secretion of tears just about keeps pace with their evaporation. When, however, tear fluid is being secreted in large quantities, it is siphoned off into the nose through the lachrymal ducts ; it may also pass between the lids but is normally prevented from doing so by the greasy secretion from their edges. The two puncta by which the tear fluids are drained into the lachrymal sac and duct can be seen on the edges of the lids on their medial aspect.

The *sclera*, as its name implies, is a tough structure serving as a protective coat to the eye. The contents of the eyeball are at about 25 mm. Hg above atmospheric pressure, so that if a cut is made in the sclera, they tend to escape. Normally the sclera resists this intra-ocular pressure, but occasionally it weakens locally and becomes bulged under the strain. Certain blood vessels and nerves pierce the sclera, the fibres of the optic nerve passing through a lattice-work of fibrous tissue known as the *lamina cribrosa*.

The sclera is lined on the inside by the *choroid*, a highly vascular coat ; lining the choroid is the retina, the light-sensitive structure of the eye. The choroid not only contains the principal blood supply to the eye, but, since it is very heavily pigmented, it prevents light both from reaching the retina through the sclera and from scattering within the eye ; a camera, similarly, consists of a light-proof box, blackened inside. The structure of the sensitive portion of the retina merits more detailed examination and will be considered later. In the region immediately posterior to the corneo-scleral junction, the retina and pigment epithelium become modified and enlarged to form the *ciliary body* (Fig. 18. 2), which consists of a stroma, a very rich blood supply, some glandular tissue and a group of muscles (the ciliary muscle). From the ciliary body there arises the *iris*, a structure composed morphologically of two layers, one continuous with the choroid, the other with the retina. The choroid layer, facing outwards and visible through the cornea, usually develops a heavy pigmentation during the period of growth. In some people this pigment fails to develop, the retinal coat can be

seen and the iris appears blue. The failure of the brown pigment to develop on the anterior surface of the iris is due to a Mendelian recessive character, the presence of the pigment being dominant. Only if both parents are without the pigment, and thus have blue eyes, are all the children similarly without it.

By far the greater bulk of the interior of the eyeball is occupied by the *intra-ocular fluid* (also known as the *aqueous humour*) and the *vitreous body* (or *vitreous humour*) separated by the *crystalline lens*. These three are transparent and, with the cornea, make up the optical system of the eye. That part of the eye which lies in front of the lens contains intra-ocular fluid only. The rest of the posterior chamber, comprising about two-thirds of the bulk of the eyeball, is occupied by the vitreous body.

In many respects the composition of the intra-ocular fluid is consistent with its being formed by a process of ultra-filtration from the

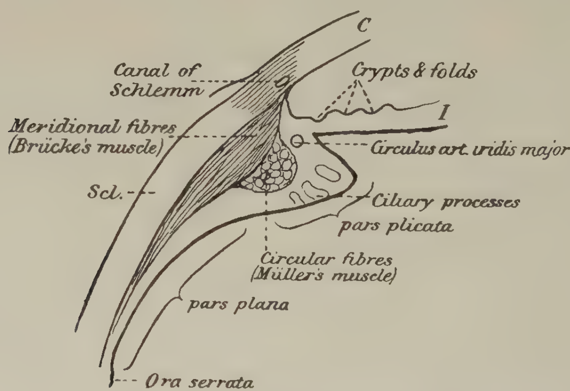


FIG. 18. 2. The Ciliary Body—an enlarged section of the region " b " in FIG. 18. 1. C = cornea ; I = iris. (Parsons.)

blood ; the membrane through which it is filtered being almost impermeable to proteins, and less permeable to large non-electrolyte molecules than to small. But accurate measurements have shown that the total (osmolar) concentration and the composition are slightly, but significantly, different from those of an ultra-filtrate, and a secretory process must be concerned in its formation. The intra-ocular fluid arises chiefly by filtration from the ciliary blood vessels, together with some secretory activity by cells of the ciliary body, and escapes partly by the " aqueous veins " which communicate with the conjunctival veins, and partly through the spaces of Fontana—minute openings in the sclera immediately posterior to the cornea—to enter the canal of Schlemm, and then the venous blood. The rate of flow is such that about 1 per cent. of the volume of the fluid is changed every minute. If a considerable fraction of the amount normally present is withdrawn, it will be replaced in about one hour : the new fluid has the composition of an ultra-filtrate of plasma, and the changes due to secretory activity are only brought about very slowly.

The vitreous body is a thin transparent jelly of about the consistency of the white of a raw egg, and containing 99 per cent. water. It has no obvious microscopic structure, although, like most gels, it consists of a felt-work of ultra-microscopic fibres. The gel is formed chiefly by the polysaccharide, hyaluronic acid, although it also contains protein (the " residual protein "). The vitreous is a much more stable constituent of the eye than the aqueous and losses may have to be

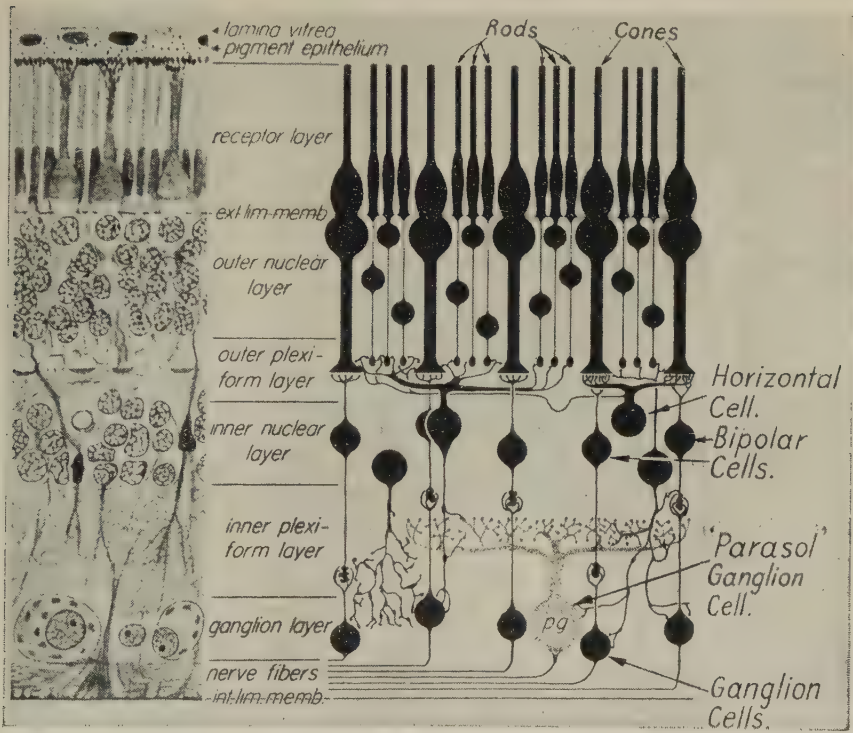


FIG. 18. 3. The Human Retina.

At the left, a vertica section through the retina in the nasal fundus, as it appears n ordinary histological preparations ($\times$ 500). Note cross-section of capillary in inner nuclear layer.

At the right, a " wiring diagram " of the retina showing examples of its principal elements. (After Walls.)

compensated by a serous exudate behind the retina often leading to retinal detachments.

Microscopically the *retina* (Fig. 18. 3) is composed of several layers, the outermost of which contains the light-sensitive receptor cells known as *rods* and *cones* on account of their shape. The nuclei of these form the outer nuclear layer. There are two other cell layers, the inner nuclear and ganglion cell layers, the latter giving rise to the axons of the optic nerve. The arborisations of the nerve fibres connecting the cell layers form two intermediate zones called the outer and inner fibre

(plexiform or molecular) layers. Detailed histological study of the organisation of these layers has revealed that the interconnections between the receptor cells, the cells of the inner nuclear layer, and the ganglion cells are extremely complex ; their nature, however, is of great importance in interpreting the phenomena of vision in terms of the probable properties of the receptors themselves. The first important point to notice is that many rods are always connected to a single bipolar cell, which has widely spreading dendrites (these cells are known as the " diffuse," and " mop " bipolars). Each bipolar cell, in turn, connects with several ganglion cells. On the other hand, each cone is connected with a single " midget " bipolar cell, and thence to a single ganglion cell. There must be a considerable summation of the activities of many rods, whereas each cone may have a " private line " to the brain. The cones, however, also have connections with " diffuse " bipolar cells, so that the activities of several may be combined with each other, and with those of the rods : the " horizontal " cells, also, appear to connect together several cones and groups of rods, which may be at some distance from one another. Lastly, there are large " parasol " ganglion cells, whose dendrites spread widely, and connect with large numbers of bipolar cells of all kinds, distributed over quite a wide area ; and " amacrine " cells (not labelled in Fig. 18. 3) which appear to connect together a number of ganglion cells. In addition there are cells whose connections appear to be directed in the wrong way, from the ganglion cells to the receptor cells. It must be remembered, of course, that these connections need not always be functionally active, and that they may be either excitatory or inhibitory. Their significance is by no means understood.

It is a remarkable thing that the rays of light, before they can reach the actual sensitive layer of receptors, have to pass, not only through the inner layers of the retina, but also betweeen the branches of the retinal vein and artery as they ramify on its surface. Although these structures are fairly transparent, their presence in the path of the beam must cause some interference with the perfection of the image in man and animals having this type of retina.

This arrangement is apparently a developmental accident, so to speak, related to the particular way in which the neural tube is formed, in vertebrates, by an invagination of the ectoderm. The eyes of the molluscs which, particularly in the cephalopods, may be as elaborate and effective as those of many vertebrates, have the sensitive elements of the optic nerve endings pointing towards the light, and away from the outer coverings of the eyeball.

The presence of the retinal blood vessels can be appreciated in one's own eye under certain conditions. One of the best ways of doing this is to illuminate a small area of the sclera, well away from the cornea, very brightly from outside the eye. This illuminated area casts a shadow of the vessels on the sensitive layer, which is interpreted as an image of objects in the external world ; one sees a large branching pattern on one's field of vision. On moving the light the vessels appear to move. By measuring the actual movement of the light, and the apparent movement of the vessels, it is possible to calculate the distance between the vessels and the sensitive layer. Since this calculated

distance is found to be the same as that between the vessels and the rod and cone layer, it follows that the latter must contain the visual receptor cells.

At the posterior pole of the eye there is a small depression on the retinal surface called the *fovea centralis* (central pit) (Fig. 18. 4). When we look at a small object, the eye is moved so that the image of that object falls on the fovea. It is a late development in mammals, and its evolution can be followed in the primates, the lower members of the series having none, whilst it becomes gradually more marked, reaching its highest development in man. The structure of the fovea is such that it allows more distinct vision than any other part of the retina. It is encircled by blood vessels but itself has none, and the nuclear and fibre layers of the retina are swept to one side ; the path of the light rays is thus uninterrupted. The fovea lies in the centre of the macular plexus,

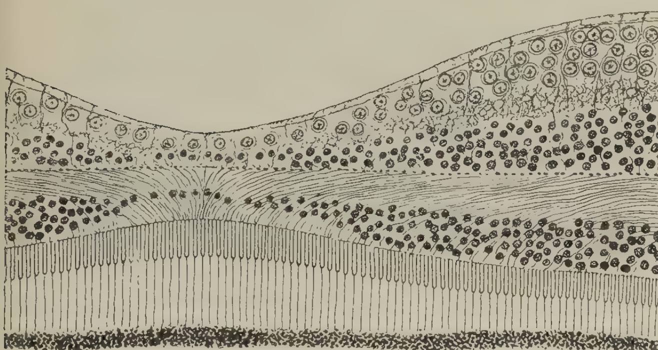


FIG. 18. 4. A cross-section of the **Retina** showing the **Fovea Centralis.**
An enlarged section of the region "a" in Fig. 18. 1. (After Sobotta.)
There are no rods in the retina at this point.

a region of the retina where the ganglion cell layer and the outer fibre layer are much proliferated.

Corresponding with the macular plexus is a thin layer of pigment on the surface of the retina known as the macula lutea (yellow spot). Its function is unknown ; its presence is troublesome in colour vision experiments, since it interferes with colour matching, and its amount varies in different people. Its presence can be appreciated subjectively by opening and closing one eye in front of which is a bottle containing a green solution of chrome alum ; on looking at the sky the macula lutea can be seen as a pink spot, subtending about 5°, lying on the green ground formed by the rest of the sky.[1]

The Formation of the Retinal Image

It is important to remember that the strongly curved outer surface of the cornea is the chief refracting surface of the eye. The lens is very powerful if examined in air, but since it is suspended in fluids whose refractive indices are only a little less than its own, it loses most of its

[1] With a clenched fist, the distance between the second and fifth knuckles subtends an angle of 8° at the eye when the hand is held at arm's length. The distance between the second and third knuckles subtends 3°.

power when in the eye. The lens is concerned with the fine focusing of the image on the retina. Application of the laws of optics to the refracting surfaces and the refractive indices of the eye media shows that the retinal image must be inverted, just as is that formed by the lens of a camera. Such an inverted image can be seen in the excised eye of an albino rabbit by looking at it from behind through the non-pigmented sclera. In infancy, during the period when the sense of vision is being developed, we learn to associate the inverted image with the uninverted external object. If from any cause in later life an uninverted image is formed on the retina, the object itself appears inverted.

This can be demonstrated by casting the shadow of an object, held very close to the eye, on to the retina by means of a point source of light. If the object is a pin head upwards, it is seen head downwards. For a similar reason, when an effusion of blood has made its way into the vitreous and the clot sinks from the top to the bottom of the eye, the shadow of the clot appears to pass from below upwards. Additional evidence of an inverted retinal image is derived from the relative positions of the *fovea centralis* and the place of entry of the optic nerve (blind spot). Also, in diseases leading to blindness of one-half of the retina, objects in the outer world lying on the side opposite to the lesion are not seen.

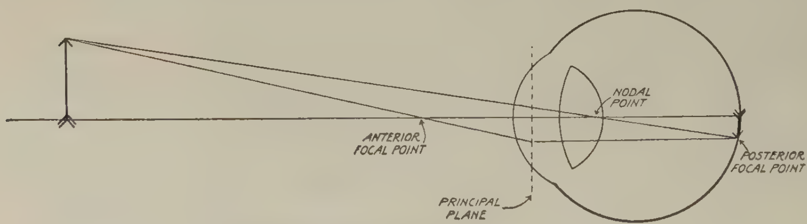


FIG. 18. 5. The **Formation of the Retinal Image** in the Schematic Eye.

Before reaching the retina the rays of light have to pass through the cornea, the intra-ocular fluid, the lens and the vitreous body. The absorption of visible rays by these eye media is small except in old age, when there is an appreciable absorption of the shorter spectral waves by the lens. Infra-red and ultra-violet radiations, both of which are capable of damaging the retina, are absorbed by the eye media before reaching it. Both these types of radiation are harmful to other structures in the eye. The infra-red rays, for instance, can cause cataract of the crystalline lens (*e.g.* amongst glass-blowers), whilst the ultra-violet rays cause intense inflammation of the conjunctiva (film-star's eye, snow-blindness). The retina is protected against intense visible radiation by changes in the pupil diameter, the iris contracting reflexly under these conditions and dilating again more slowly in the dark. Even when the pupil is fully contracted, the absorption of visible light by the retina at the site of a very bright image may lead to a rise of temperature sufficient to cause a burn resulting in local blindness. This condition is not uncommon in those who unwisely observe an eclipse of the sun without adequate protection by dark glasses.

The image on the retina is formed by what, in geometrical optics, is

called a " thick lens." The size and position of the image can thus be calculated from a knowledge of the cardinal points of the optical system. The most useful of these is the *nodal point*, an imaginary point at the optical centre : it lies at the geometrical centre of a thin lens, but in the eye it is in the crystalline lens near the posterior surface, 15 mm. in front of the retina (Fig. 18. 5). A pencil of light directed towards the nodal point behaves as though it had passed through undeflected, although this does not mean that the pencil actually follows this path. Its use is in calculating the size of the retinal image, when the size and distance of the external object are known. The calculation is made by the use of the principle of similar triangles. It can be applied, for instance, to the calculation of the distance between the blind-spot (optic disc) and the region of most distinct vision (fovea centralis). A circle and, 60 mm. to its left, a cross are

Fig. 18. 6. Figure for illustrating the Presence of the **Blind-spot.**
Look at the cross with the right eye only and hold the book at about 22 cm. from the eye.

made on a piece of paper (Fig. 18. 6), and, looking at the cross with the right eye only, the paper is moved away from the eye. At about 220 mm. distance the circle will be invisible because its image will fall on the blind-spot, at which point the sensitive elements in the retina are absent. By " similar triangles " we have :—retinal distance required : 15 mm. : : 60 mm. : 220 mm. The distance between the fovea and the centre of the optic disc works out to be 4 mm., a result which agrees with histological observation.

The *anterior focal point* is 13 mm. in front of the cornea, and a pencil of rays which has passed through it is refracted parallel to the optic axis when it meets the *principal plane*, which lies 2 mm. behind the anterior surface of the cornea (Fig. 18. 5). Using these quantities in conjunction with the nodal point, the position of the image within the eye can be determined.

The eye behaves like other optical instruments in that it is subject to some of their defects. Among these can be mentioned *spherical aberration*. The power of a spherical lens is greater at the periphery than at the centre, that is to say, the image formed by its periphery is nearer to the lens than the image formed by its centre. In the eye the outline of the refracting surfaces is slightly hyperbolic and in this way

spherical aberration is diminished. *Chromatic aberration* of a lens system is due to the fact that the short (blue) waves of the spectrum are bent at a refracting surface through a greater angle than are the long (red) waves, and so come to an earlier focus. Actually the eye, on looking at a point source of light, is focused on the middle wave-lengths, giving the light a halo of red and blue. These halos can be seen by looking at a small light through a piece of cobalt glass which transmits only blue and red rays.

Accommodation. A normal eye, directed on distant objects, will bring their images to a focus on the retina ; its *far point* of distinct vision is said to be at infinity. If an object is brought close to the eye in these conditions, the image becomes blurred, just as in photography the image of a near object becomes blurred unless suitable adjustment of the lens system is made. A corresponding adjustment is made by the eye and is known as *accommodation.* There is a limit beyond which the eye cannot further accommodate, and the nearest point at

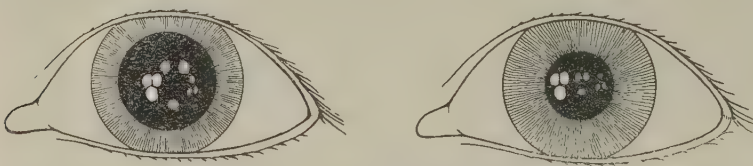


FIG. 18. 7. The Changes in the **Purkinje Images** during Accommodation.
A 3-point source of light forms mirror images at the anterior surface of the cornea and at the anterior and posterior lens surfaces. During accommodation (right) only the anterior lens image changes. Notice the contraction of the pupil. (Fincham.)

which objects can be seen distinctly after maximum accommodation is known as the *near point.* The mechanism of accommodation by which the divergent rays from near objects are converged to a focus on the retina is not the same in all animals. In mammals, including man, the lens is not moved, as it is in a camera, but its power is increased by an increase in its curvature.

The changes that occur in the lens of the human eye can be observed by means of the Purkinje images (Sanson's images). These are images of a source of light, held slightly to one side of the subject's eye, formed by reflection at the principal refracting surfaces. The anterior surface of the cornea acts like a convex mirror and forms a bright uninverted image. The anterior surface of the lens functions similarly, but the image is faint and often very difficult to see, and normally is larger than that formed by the cornea. The posterior surface of the lens acts as a concave mirror ; the image is inverted and much smaller than the others. When the eye accommodates alternately for far and near objects a marked change occurs in the image from the anterior lens surface ; there is a very slight change in the image from the posterior lens surface, and no change at all in the image from the cornea, showing that this structure is unchanged during accommodation (Fig. 18. 7). In order

to focus near objects the curvature of the anterior lens surface is increased and consequently the mirror image formed at its surface becomes smaller. When the eye is focused on infinity, the radius of curvature of the anterior surface of the lens is about 10 to 12 mm. ; during maximum accommodation, the radius of curvature is reduced to about 6 mm. The radius of curvature of the posterior surface changes only slightly, from about 6 mm. to about 5·5 mm. It can be shown by direct microscopical observation, also, that the anterior lens surface moves forward during accommodation, while the posterior surface moves backwards very slightly, the position of the lens as a whole being almost unchanged.

An excised lens has surfaces as highly curved as in the eye when accommodated for near vision. The reason for this is that the lens is an elastic body which, for distant vision, is stretched and flattened as a result of the method by which it is supported ; and for near vision

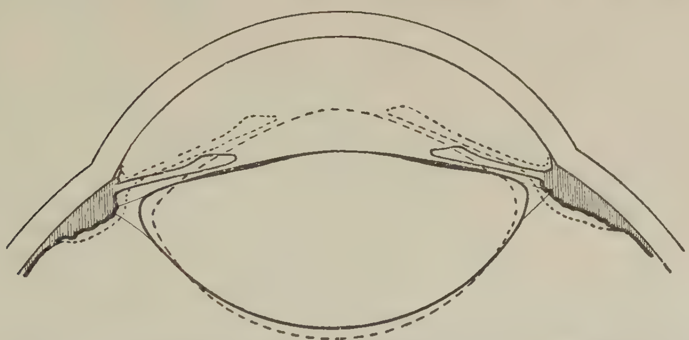


FIG. 18. 8. Diagrammatic Cross-section showing the **Changes during Accommodation** (dotted lines). (After Helmholtz and Fincham.)
Note the regional variation in the thickness of the lens capsule.

is released and allowed to thicken. The support for the lens is provided by the *suspensory ligament*. This is composed of elastic strands attached to the periphery of the lens capsule and they radiate outwards to become attached to the ciliary body. This body contains plain muscle fibres, which run more or less circumferentially near the inner margin, and more or less radially near their attachments to the choroid. They contract when the eye is accommodated for near vision, and the inner edge of the ciliary body is drawn inwards, towards the lens, and slightly forwards, towards the cornea ; this movement results in a reduction in the pull exerted on the lens by the suspensory ligament. The lens substance is held under pressure exerted by the stretched capsule, and may be observed to flow out when a small cut is made. The capsule, however, is not of uniform thickness : when the lens is stretched or released, the surface curvature will change most in those regions where the capsule is thinnest, that is the central part of the anterior surface. The more peripheral part of the anterior surface, where the capsule is relatively thick, does not alter much in curvature : but during accommodation the pupil constricts, so cutting off the peripheral parts of the

lens, allowing only the most curved region to take part in the formation of the retinal image.

The changes during accommodation can be summarised into (1), contraction of the ciliary muscles, (2) the approximation of the ciliary body to the lens, (3) relaxation of the tension of the suspensory ligament, (4) increased curvature (bulging) of the anterior surface of the lens due to the tension in the capsule, and its thinness in the central region (Fig. 18. 8). In addition, during accommodation, the pupil contracts and the optic axes converge. These three reflexes are very closely associated with one another and they are often grouped together as the *near reflex* or *accommodation reflex*. All three are mediated by the IIIrd (oculomotor) nerve and are designed to facilitate the production of a sharp image on corresponding points of the two retinæ. The relationship between accommodation and convergence will be considered later.

The IIIrd nerve, as discussed in Chapter 15, forms part of the parasympathetic nervous system, and its nerve endings are cholinergic. Atropine, homatropine, and allied drugs thus abolish the power of accommodation, and are used clinically for this purpose. Since they keep the intra-ocular muscles at rest, they are also used in the treatment of intra-ocular inflammations. The effect of one drop of 1 per cent. homatropine on the eye is maximal in about an hour and takes about twenty-four hours to wear off completely. The effect of atropine lasts nearly a week. Eserine (physostigmine) and pilocarpine have opposite effects ; they cause a spasm of the ciliary muscles and the eye is focused for near objects. The action of atropine is partially overcome by these drugs and they are sometimes used to make reading possible after an examination of the eye. Eserine is interesting because it renders the suspensory ligament of the lens completely slack and the lens itself can often be seen to lie loose in the eye.

The **power of accommodation** is measured in the following way. The subject is given a printed page and told to hold it at the near point, that is, with maximum accommodation. He is then told to relax the accommodation and convex lenses are placed in front of the eyes until the print, held in the same position, is again clearly seen. The power of this lens is a measure by which the lens in the eye had to increase its power in order to see the print clearly in the first place.

In ophthalmology the power of a lens is expressed, not by its focal length, but in *dioptres* (D), that is, the reciprocal of the focal length in metres. A convex (converging) lens with a focal length of 2 metres is a + 0·5 D lens ; a concave (diverging) lens of the same focal length is a − 0·5 D lens. The advantage of this system is that the powers of lenses are additive. For example, the combination of a concave lens of − 4 D (focal length 0·25 m.) with a convex lens of + 5 D (focal length 0·2 m.) is equivalent to a convex lens of + 1 D (focal length 1·0 m.). Thus in the above example, if the subject held the printed page 0·2 metres from his eyes, the auxiliary lenses would have a focal length of 0·2 m., or a power of + 5 D ; they would make the light entering the eye parallel, and the page would appear to be at infinity.

Accordingly, when the subject used maximum accommodation, he increased the power of his own lenses by 5 D. This is called exercising five dioptres of accommodation.

The ability to accommodate becomes smaller with increasing age owing to a progressive hardening of the lens, and the near point gets further away. The condition is known as *presbyopia* (old sight). Usually the change is small up to the age of about forty, but in older people it comes on increasingly rapidly, as shown in Table 18. 1 ; at the age of eighty, all power of accommodation is usually lost. Presbyopia is treated by giving the patient convex spectacles for near work ; they give the incident rays an initial convergence and compensate for the inadequate convergence given by the crystalline lens. Such spectacles become essential for reading when the near point has receded beyond arm's length—*i.e.* usually at the age of about fifty. But some people need them before this age, and a few can still read without spectacles at the age of eighty.

TABLE 18. 1

The Onset of Presbyopia
(*Approximate figures for normal eyes*)

Age, years	10	25	35	45	50	60	70	80
Near Point, cm.	8	10	15	28	55	80	100	∞
Power of Accommodation, dioptres	12·5	10	6·7	3·5	1·8	1·25	1	0

Errors of Refraction. It is unusual amongst either man or other animals to find an eye which is " normal " in respect to its optical system. Two of these optical defects (*errors of refraction*) are very common : *hypermetropia* (long sight), in which the focus of the unaccommodated eye is behind the retina, and *myopia* (short sight), in which the focus of the unaccommodated eye lies in the vitreous (Fig. 18. 9). The eye is always hypermetropic at birth, since the lens is then almost full sized and grows very little in after life ; the growth of the rest of the eyeball follows more closely that of the body generally. At birth, therefore, the eyeball is too short relative to the power of the lens and the rays fall on the retina before coming to a focus. With the subsequent growth in after years this defect is normally remedied, but in some people the growth of the eyeball is arrested and the eye remains hypermetropic. Myopia is almost always due to an excessive lengthening of the eyeball, a fact which can be verified by asking a myope to look as far as possible towards his nose ; so exposing an unusually large area of sclera.

People suffering from hypermetropia can usually remedy the defect by accommodating ; indeed they must continually use accommodation to see anything, including distant objects, clearly. " Eye-strain " is

common in this condition, and is possibly due to constant use of the ciliary muscle. Myopes can never see distant objects clearly unless the incident rays are slightly diverged by concave spectacles (Fig. 18. 9). Their far points, at which objects are focused accurately without accommodation, are not at infinity, but are relatively close to them. Hypermetropes cannot see objects as close to the eye as normal people can, since they cannot produce the necessary amount of accommodation ; their near points are farther off. Myopes, on the other hand, have a near point which is closer to the eye than normal. Hypermetropes, since they begin by having a too-distant near point, and since this recedes

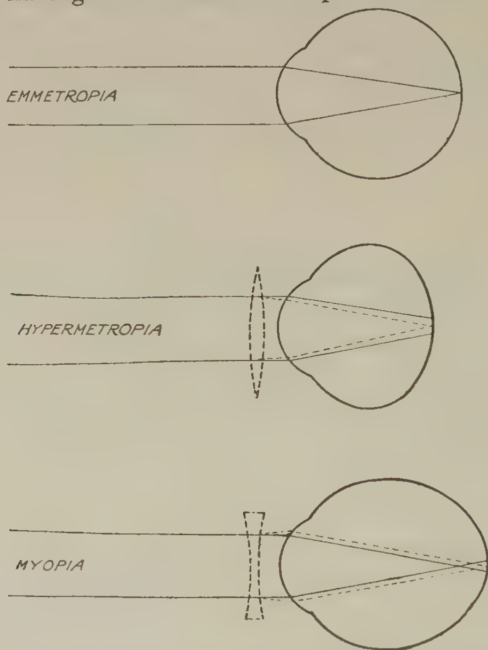


FIG. 18. 9. **Errors of Refraction.** Diagram of the paths of parallel rays on entering the Emmetropic (normal), Hypermetropic (long-sighted) and Myopic (short-sighted) eyes. The dotted lines show the paths taken when a correcting lens is introduced.

progressively with age, have to take to reading glasses earlier. A myope may never need reading glasses, since his near point only recedes as far as his far point, which may be within reading distance. Degrees of myopia and hypermetropia are usually expressed in terms of the number of dioptres necessary to give normal vision ; thus, a myopic eye which needs a concave lens of -3 D would be said to have three dioptres of myopia.

In myopia and hypermetropia, the lens and its changes during accommodation are normal. The increase in the power of the lens during accommodation, as measured in the way described above, is the same as that of people with normal vision, provided, of course, that their ages are about the same. If a hypermetrope, for example, who ordinarily

has to accommodate by, say, + 1·5 D in order to see clearly in the distance, is to focus on an object 0·2 m. away, he must increase his accommodation by the same amount as a normal person, *i.e.* by + 5 D. He must thus exercise + 6·5 D of accommodation altogether, and his near vision is likely to fail, with increasing age, sooner than that of a normal person.

One other common optical defect of the eye must be mentioned, namely *astigmatism.* This is a condition where the foci for vertical and horizontal rays entering the eye are different (Fig. 18. 10). Astigmatism is usually due to differences in the curvature of the cornea in two directions, although astigmatic lenses occur, too. If, for instance, the curvature of the cornea were normal horizontally but too great vertically, then with a horizontal slit in front of the eye the rays would all come to a focus on the retina, but with a vertical slit they would be focused in front of the retina.

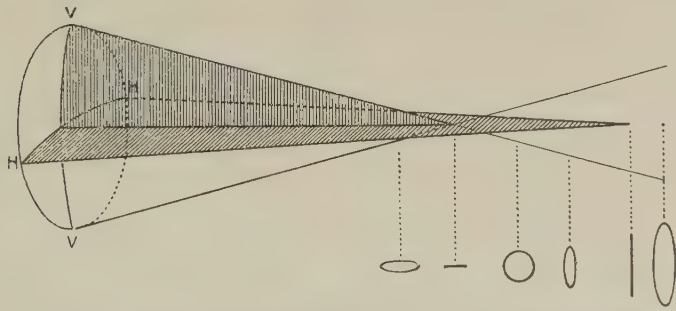


Fig. 18. 10. Diagram showing the Paths of Rays in Astigmatism.
The vertical meridian (V, V) of the refracting surface is curved more than the horizontal meridian (H, H). Cross-sections of the conoid of light are shown below.

The Pupil. The iris can be looked upon as a diaphragm capable of varying the size of its aperture in response to different stimuli. The aperture is called the *pupil* and through it light enters the eye. The contraction of the pupil not only prevents the eye from receiving too much light but also improves the definition of objects.

The iris has two muscles, the circular muscle (*sphincter pupillæ*) contraction of which constricts the pupil, and the radial muscle (*dilator pupillæ*) stimulation of which causes dilatation of the pupil. The circular muscle is supplied by the IIIrd nerve and the radial by the sympathetic. Both sets of muscle fibres are also in a state of tone which is independent of their nervous connections, but the circular are stronger than the radial fibres, and normally are able to overcome them. In ordinary circumstances, the iris is in a constant state of activity governed by the antagonism between these two muscles. Cutting the IIIrd nerve fibres to the eye causes a moderate dilatation of the pupil while stimulation of the cut ends makes it constrict. Stimulation of the sympathetic produces dilatation.

This dual innervation is reflected in the rather complex action of

drugs on the pupil. Since the IIIrd nerve belongs to the parasympathetic system, its action on the circular muscle is inhibited by atropine and homatropine ; the normal tone of the radial muscle is left unopposed, and the pupil dilates. The pupil, moreover, can no longer constrict when a strong light is directed into the eye, so that everything appears dazzlingly bright. Conversely, the parasympathomimetic drugs, eserine and pilocarpine, stimulate the circular muscle and the pupil constricts. Furthermore, the adrenergic endings of the sympathetic nerves may be potentiated by the use of cocaine ; the radial muscle is caused to contract and so the pupil dilates. This occurs to a slight extent even after the circular muscle has been paralysed. Homatropine and cocaine make a useful combination, since not only do they together ensure a maximum dilatation of the pupil, but cocaine also anæsthetises the conjunctiva whereas homatropine is apt to make it smart. Novocaine is ineffective in the eye, since it does not penetrate either the cornea or the conjunctiva. Morphine constricts the pupil to the size of a pin's head ; the action is on the nucleus of the IIIrd nerve, and cutting this nerve abolishes the effect.

The size of the pupil is constantly changing in response to various external stimuli, the most important of which is light. An increase in the amount of light entering the eye causes constriction of the pupil while a decrease causes dilatation. Most of this reaction is due to a reflex following stimulation of the retina, but it is probable that light also has a direct effect on the iris muscle in man. In all animals in which there is partial decussation of the optic nerve (see Chapter 14), illumination of one retina produces constriction of the pupil in the unstimulated as well as the stimulated eye. This is known as the *consensual reflex.*

The effectiveness of light in eliciting pupillary constriction depends upon *changes* of illumination rather than upon the absolute quantity of light entering the eye. Thus, a rapid increase in illumination will produce a much greater initial constriction than a slower increase over the same range. Also, if two people are brought under the same moderate illumination, one after having been in the dark and the other after subjection to a high illumination, the pupils of the first will constrict and those of the second will dilate. Later, as the eyes of both become adapted to the new illumination, the pupils of the first subject will dilate and those of the second constrict until, for both, they have reached the physiological size for those particular conditions of illumination. This property of responding by pupillary constriction to a higher illumination, and dilatation to a lower illumination, than that to which the eye happens to be attuned at the time is bound up with the general phenomenon of adaptation to different conditions of illumination. This whole subject of adaptation will be dealt with in detail later.

The Sensation of Vision in Man

Day Vision and Night Vision. The properties of human vision are markedly different according as the intensity of illumination is below or

above a certain value (about 0·01 foot-candles or 0·1 lux), which is approximately that of bright moonlight. The differences are sufficiently great for night vision (or **scotopic** vision) to be sharply distinguished from day vision (or **photopic** vision). At high intensities of illumination, we can appreciate colour (differences in wave-length of the light) and fine details ; the brightest part of the spectrum is in the yellow, at a wave-length of about 560 mμ (Fig. 18. 11). At low intensities of illumination, we are colour-blind ; the whole spectrum is appreciated in shades of grey, the brightest part being at a wave-length of 500 to 505 mμ (green if colours could be appreciated), and the " red " end of the spectrum, at wave-lengths greater than 650 mμ, being hardly seen at

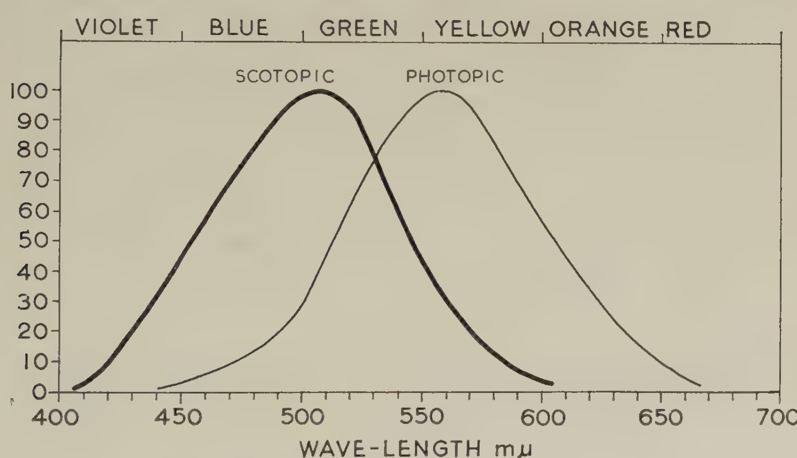


FIG. 18. 11. Scotopic and Photopic **Luminosity Curves.**

Abscissæ—wave-length in mμ.

Ordinates—relative sensitivity of the eye, the sensitivity at the most effective wave-length (505 mμ for scotopic conditions, and 560 mμ for photopic conditions) being made, arbitrarily, equal to 100. (Crawford, and L. C. Thomson.)

all. These changes in the relative brightness of colours with intensity of illumination are known collectively as the *Purkinje shift.*

If, however, we confine our attention to the fovea, we find that nothing is seen unless the intensity is great enough for colour and detail to be appreciated. At the extreme periphery of the retina, on the other hand, colour and detail are poorly appreciated, even in moderately high illuminations. The fovea is capable of day vision only, while night vision is the concern of the peripheral parts of the retina. At night, a faint light, such as a feeble star or dimly illuminated window in the distance, may be visible if not looked at directly, but disappear when fixated, and the image brought on to the fovea.

Microscopic study of the retina shows that the fovea is composed entirely of cones while around it they become mixed with rods ; and, in general, as one goes farther out towards the periphery the relative

number of rods to cones becomes progressively greater. We may thus connect the cones with the acute coloured vision characteristic of high illuminations, mediated by the fovea; the rods, which predominate in the periphery, with the colourless, but more sensitive vision of low illuminations. This association of cones with day vision and of rods with night vision is known as the *Duplicity Theory.*

The cones are sensitive to intensities of illumination much smaller than that at which day vision ceases (see Fig. 18. 12, below); and rods may still function when the brightness is such that night vision is replaced by day vision. The duplicity theory does not deny this, but states that in dim light, the responses from the cones are negligible compared with those from the rods, whereas in bright light, the cone responses swamp the rod responses.

The duplicity theory is borne out by the fact that, in general, animals that live mostly by night or in dark places have a great predominance of rods in their retinæ, while animals which spend much time in the sun have mostly cones. This relationship between the rod and cone content of the retina and the habit of a species is particularly well developed in certain day- and night-living species of the lizard family, as well as in the primates.

Some people suffer from a condition known as " night-blindness "; their day vision is normal, but their night vision, in low intensities of illumination, is deficient. In some, this is due to the absence of rods from the retina, resulting either from a congenital defect (which may be inherited), or from their destruction by disease, as in retinitis pigmentosa. In others, the deficiency is due to an inadequate production of the photochemical substance visual purple, which will be discussed later.

Dark-Adaptation. It is a matter of common experience that the eye works well even though the intensity of illumination varies over a very wide range. Some process of adaptation occurs, so long as the eyes are being used, adjusting the sensitivity of the eye—and, indeed, many of its other properties as well—according to the prevailing brightness of the objects in the field of view. The term " adaptation " is used to describe either the process of adjustment, or the final " stationary state " after adjustment is complete. The state of adaptation, at any moment, is measured in terms of the *visual threshold,* that is, of the smallest brightness just necessary to stimulate the eye. It is not possible, however, to find a value of the brightness which is always seen, and another slightly smaller value, which is never seen: the threshold value, therefore, is commonly defined as the brightness of a test object which is seen on one-half of the occasions on which it is presented. Clearly, if one wishes to maintain a steady condition of adaptation, one must keep the eyes in light of constant intensity. The sensitivity is then measured in terms of the least increase (or decrease) in intensity which can just be detected: this is known as the *increment* (or *difference*) threshold. If the constant adapting light has an intensity I, and the difference threshold is Δ I, then Δ I increases as I increases; the sensitivity becomes less when the eyes are adapted to brighter lights. Over a wide range of moderately bright illumination, the fractional difference threshold, Δ I/I, is found to be fairly constant, at a value of about 0·03. When the background intensity is small, however, the fractional difference threshold increases as the intensity is reduced,

although the difference threshold itself continues to become smaller. This may be seen from the values given in Fig. 18. 13 (p. 540, below).

If one passes suddenly from brightly illuminated surroundings to, say, a room which is only dimly lit, the adaptation of the eye is found to be a slow process, taking up to an hour or more ; but the change in sensitivity is very great, as is indicated in Fig. 18. 12. The actual increase in sensitivity depends on the illumination to which the eye has previously been subjected, but the fully dark-adapted eye can appreciate a light of only one hundred-thousandth of the intensity

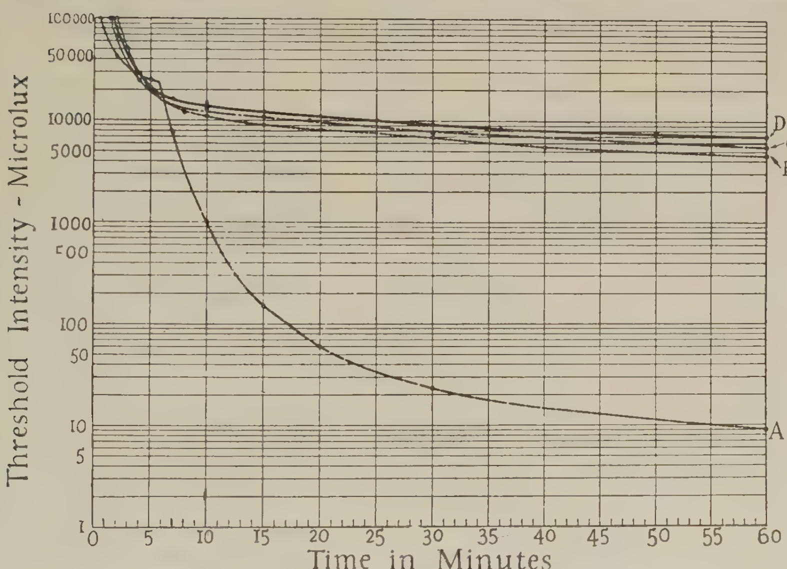


Fig. 18. 12. Dark-adaptation Curves of one normal (A) and three congenitally night-blind (B, C and D) subjects.
1 lux = 1 metre-candle = 0·1 foot candle (very nearly). (After Dieter.)

necessary to stimulate it when it is light-adapted. When fully dark-adapted, an average person can just see (on half the occasions that he looks at it) a large white surface which has an illumination of about 10^{-6} cd/m^2; this would be produced by an ordinary candle which is about 600 yards away. This is the *absolute threshold* of vision. Among quite normal individuals, however, there is a 10-fold range of variation in this figure. In bright sunlight, the illumination on the white surface would be about 3×10^{10} times greater than the absolute threshold.

On returning to the brightly lit surroundings, the sensitivity falls again very rapidly, by a factor of 1,000 or so, in less than a second. But if one stays in the light for only a few seconds and then returns to the dark room, full sensitivity is regained quite rapidly, in no more than a few minutes. The complete adaptation to bright illumination, which is replaced by dark-adaptation only after an hour or so, as illustrated in Fig. 18. 12, occurs only if one stays in the light for several minutes.

The measurement of dark-adaptation has been the subject of a vast amount of work and many instruments have been designed for the purpose, but the basic principle of all methods is the same. After a preliminary subjection to light of high intensity the subject is put into the dark and the lowest intensity of light which he can recognise is measured at various intervals up to about an hour after the removal of the bright light. The results are usually presented in the form of a dark-adaptation curve (Fig. 18. 12). It will be seen that the normal dark-adaptation curve is not continuous but is divided into two quite distinct parts. It is really made up of two separate curves superimposed, the upper one being due to the adaptation of the cones and the lower to that of the rods.

The dark-adaptation curve of the cones can be obtained by using a red light of wave-length too long to stimulate the rods. Similar curves are obtained, also, from persons suffering fron congenital night-blindness (Curves B, C, D), confirming the fact that their rod vision is deficient. Thus, although the cones are also able to increase their sensitivity at low illuminations, the extent of this increase is very small compared with that of the rods.

A decrease in the illumination of the surroundings will be accompanied by an increase in the size of the pupil, as already mentioned ; there will be a smaller decrease in the illumination of the retina, or perhaps, even, no decrease at all. This effect plays some part in the phenomenon of dark-adaptation, particularly in the more rapid changes in the sensitivity of the eye, but it can only be a minor part. A change from a state of full constriction of the pupil to a state of full dilatation will increase the sensitivity by a factor of about 12, only a small fraction of the whole change in sensitivity ; dark-adaptation occurs, also, even when the effects of changes in pupil size are eliminated. The greater part of the process of dark-adaptation must be due to changes in the receptor cells in the retina, and in their connections with the nervous system.

The Visual Receptive Field. If light is arranged to fall on a small and known area of the retina, it is found that the least intensity which can be seen is inversely related to the area illuminated. The fact that a decrease of intensity can be compensated by an increase in area shows that the responses of several receptor cells, each insufficient by itself, can add together. This observation supports the histological evidence described above, and the electrophysiological evidence to be discussed below, that a great many receptor cells may be connected to a single fibre in the optic nerve, and that a given receptor cell may be connected to several different nerve fibres. If, however, the area of retina illuminated is increased beyond a certain value, this summation of responses from separate receptor cells no longer occurs ; the threshold intensity becomes independent of the area illuminated. The limiting area defines the size of the "receptive field," and the receptors within it form a "receptor unit," as defined in the previous chapter. The size of the receptive field cannot be stated precisely, since it has no sharp boundary.

There is a central region in which there is complete spatial summation between different receptor cells, the threshold intensity being inversely proportional to the area illuminated—a relation which is known as "*Ricco's law*": and outside this, there is a region of partial summation, in which an increase in the area illuminated has some effect in reducing the intensity threshold, but the product of intensity and area is not constant. This is illustrated in the left-hand part of Fig. 18. 13. As is seen in this figure, also, in the dark-adapted eye (zero background illumination), the receptive field is of such a size as to subtend about 1° of arc at the nodal point of the eye: but in the partially light-adapted state, the receptive field becomes smaller, and when fully light-adapted, it may subtend only a few minutes of arc. It seems probable, therefore, that a large part of the increase in sensitivity during dark-adaptation results from an increase in the size of the receptive field, and an increased convergence of many receptor cells on to a single ganglion cell: several thousand rods may co-operate in this way. This conclusion is supported by the fact that if a very small area of the retina is illuminated, subtending only one or two minutes of arc, the threshold intensity at the fovea is found to be not greatly different from that in the extra-foveal region, even when the eye is dark-adapted; whereas, as shown in Fig. 18. 12, with large fields of view, rod vision may be several hundred times as sensitive as cone vision. The sensitivity of an individual rod, indeed, is not very much greater than that of an individual cone.

The Retinal Action Time. In a steady condition of illumination, the sensation of vision is constant (after any adaptation effects have been completed) and depends on the intensity of the illumination, or on the *rate* at which light energy is received. But if the illumination is delivered in very short flashes, of varying intensity and duration, it is found that either for a given moderate sensation of light or for a just detectable threshold sensation, a reduction in intensity can be compensated by an increase in duration. Within limits, the compensation is exact, and the sensation thus depends on the product of intensity and time, or on the *total quantity* of light energy received. The fact that a decrease in intensity may be exactly compensated by an increase in the duration of the exposure is a common feature of photochemical reactions, and is known as the "*Bunsen-Roscœ law*"; that the photographic emulsion behaves in this way—at least approximately—is taken for granted. In the eye, the limiting duration over which this law holds, and there is complete temporal summation of receptor activity, is known as the "retinal action time." As may be seen in the right-hand part of Fig. 18. 13, the time during which there is temporal summation is quite short, about 0·1 sec., becoming somewhat smaller when the eye is adapted to strong illumination. The relevance of this to the nature of the photochemical reactions concerned in vision, and to the manner in which the receptor cells are excited, will be discussed later.

Measurement of the threshold intensity of a short flash of light, whose duration is less than the retinal action time, gives a measure

of the least quantity of light energy, incident on the cornea, which can just be seen. Measurement of the optical properties of the eye media—cornea, lens and vitreous—shows that if we use the green light to which the eye is most sensitive, about one-half of the light incident on the cornea will reach the retina ; and about one-fifth of that incident on the retina will be absorbed by the rods and its energy made available for their excitation. According to the Quantum Theory of radiation,

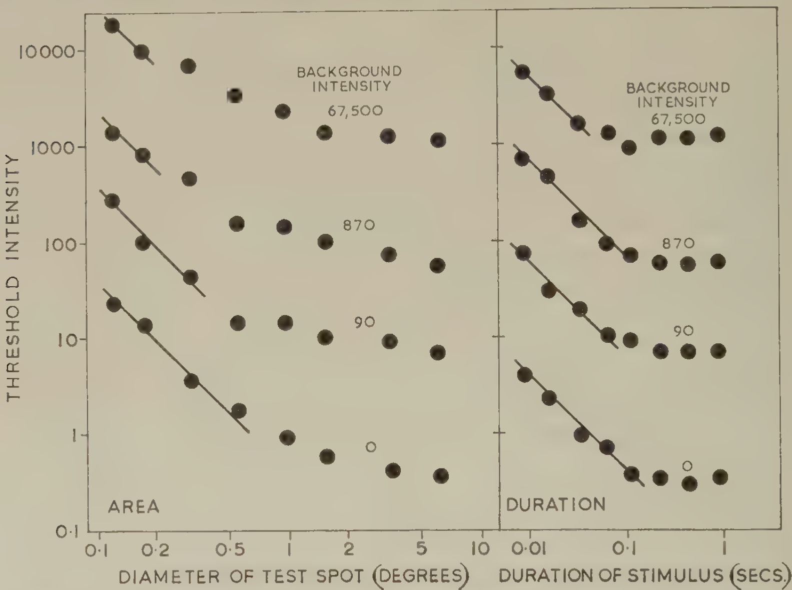


Fig. 18. 13. Spatial and Temporal Summation in the Human Eye.

The threshold intensity, or the increment threshold, is plotted against the size of the test spot (on the left), the duration of exposure being constant at 0·93 sec.; and against the duration of the test stimulus (on the right), the size of the test spot being constant, with an area subtending 27·6 deg² at the eye (diameter subtending 5·9 deg.).

The sloping lines indicate the relation to be expected if the threshold were inversely proportional to the area (Ricco's law) ; or inversely proportional to the duration (Bunsen-Roscoe law). Note the reduction in size of the receptive field—particularly of that part in which there is complete spatial summation—with increase in the intensity of the background as indicated by the figure against each group of observations. The brightest background had an illumination of about 1 lux, and was thus only moderately bright.

The intensities, both threshold and background, are expressed in units of 1,000 quanta (507 mμ) in each (degree)^2 arriving at the eye in each second. (From Barlow, redrawn.)

light energy can be emitted or absorbed only in packets of a certain size, known as *quanta* (or *photons*) ; the size of each quantum increasing as the wave-length becomes smaller. Since we know the wave-length of the light used, we can convert the threshold quantity of light energy into the threshold number of light quanta. Now in threshold conditions, by definition, the light will be seen in about 50 per cent. of the trials. But in these conditions, the number of quanta absorbed per unit area

of the retina (about 250 per square centimetre in 0·1 sec.) is very much smaller than the number of rods per unit area (about 14 million per square centimetre) : thus, according to the laws of probability it will be exceedingly rare for any single rod to absorb more than one quantum—certainly in very much less than 50 per cent. of the trials. Thus it must be possible for a rod to become activated if it absorbs a single quantum of light energy—the smallest amount which can exist.

There is no sensation of vision, however, unless several quanta—probably between 5 and 10—are absorbed within the retinal action time by a group of rods which are all connected together—*i.e.* are included in one receptor unit. This figure is based on rather indirect evidence and is still uncertain.

The Perception of Light and Shade. Visual Acuity. We obtain a great deal of information about the outer world by detecting the patterns of light and shade formed by different parts of the visual field which reflect or scatter different amounts of light. Whether these parts will be distinguished from each other or not depends, first, on their relative brightness, and secondly on the size of their images on the retina.

Two relatively large parts of the visual field will be distinguished if the fractional difference in the intensity of the light derived from them is equal to, or greater than, the fractional difference threshold, as defined above (p. 536). In general, each part of the field will reflect, or scatter, a constant fraction of the light which falls on it, the value of the fraction depending on the nature of its surface. The relative intensities of the reflected lights from the two parts of the field will thus be independent of the intensity of the general illumination. (If one part reflects 3/5 of the light falling on it, and the other 4/5, the relative intensities of the reflected lights will always be 4/3, and the fractional difference in the intensity will always be 1/3.) If the general illumination is weak, the fractional difference threshold rises as the illumination becomes still weaker : thus a grey object on a slightly paler background, for example, which is easily appreciated in bright lights, may vanish in dim lights.

The detection of fine details in the pattern of light and shade is characteristic of day vision, and of the use of the cones as receptors ; as already pointed out, it is most highly developed at the fovea, where the general arrangement is such that there is little obstruction to the path of light, and the rather specialised cones can each be connected to one fibre only in the optic nerve.

If two parts of an illuminated field of moderate brightness, are separated by a dark gap—as, for example, a white background with a black line on it—the gap can be seen provided that it subtends an angle of about one minute at the eye ; that is, the black line is appreciated, and the two fields are seen separate and not joined. Now the nodal point of the eye is about 15 mm. in front of the retina and, therefore, the width of the image of the dark gap on the retina will be about 4μ. The diameter of a foveal cone is some 2μ to 3μ, so that there will be space, and to spare, for a row of unstimulated cones to lie between

those stimulated by one field and those stimulated by the other. If the gap were much narrower, so that the width of its image was less than the diameter of a cone, there would be no unstimulated cones, and the two fields would appear to be united.

It is possible to test the perfection of an eye's optical system by this method, but the use of two fields of light is inconvenient and the letters of the alphabet (*test types*) are commonly employed for the routine testing of visual acuity. It is usually stated that a person with standard vision can recognise a letter when its details (breadth of the stroke and spaces) subtend an angle of 1 minute at the eye. This value applies only to black letters on a white ground and when the illumination is approximately that of a well-lighted room. The letters composing a sheet of test type are arranged in different sizes and, opposite each size of letter, there is a figure indicating the distance at which a person with standard vision can read that size. If a person cannot read the 18 m. line further off than 6 m., his visual acuity is expressed as a fraction of the standard, namely 6/18. It should be noticed that at 6 m., the details of the 18 m. line subtend an angle of 3 minutes and the fraction 6/18 is the reciprocal of the smallest visual angle recognisable by the patient.

The acuity of vision depends quite appreciably on the intensity of the illumination (Fig. 18. 14). The higher the illumination the greater the visual acuity ; in other words, a line of print which is too small to be read at a given distance may become easily recognisable if it is more brightly lit. This effect is of considerable practical importance. It is not to be expected from the treatment of the relation between visual acuity and the size of the receptor cells in the retina just given ; but this is because the treatment is greatly over-simplified. To begin with, for the gap between the illuminated fields to be detected, it is not necessary that the cones on which its image falls should receive no illumination at all : it is only necessary that the illumination falling on them should be less than that falling on the cones on each side of them by an amount equal to the difference threshold. The width of the image of the gap may thus be only a fraction of the diameter of a cone, and still be seen ; and the blacker the gap, the narrower it may be. Indeed, in very favourable conditions, a dark gap subtending only 0·5 second at the eye can be detected. Further, we should consider as the receptor unit, not necessarily a single cone, but the whole group of cones and rods in each receptive field. That part of the receptive field over which there is complete spatial summation will behave as a single unit, in that it will respond in the same way whatever the distribution of light and shade in its different parts. The width of the image of the dark gap, therefore, must be comparable with the diameter, not of a single cell, but of the whole receptive field : since this increases as the intensity of illumination falls, it is to be expected that the acuity will decrease, as in fact is observed. The matter, however, is more complicated than this : some of the factors concerned will be mentioned later, but not all of them are known.

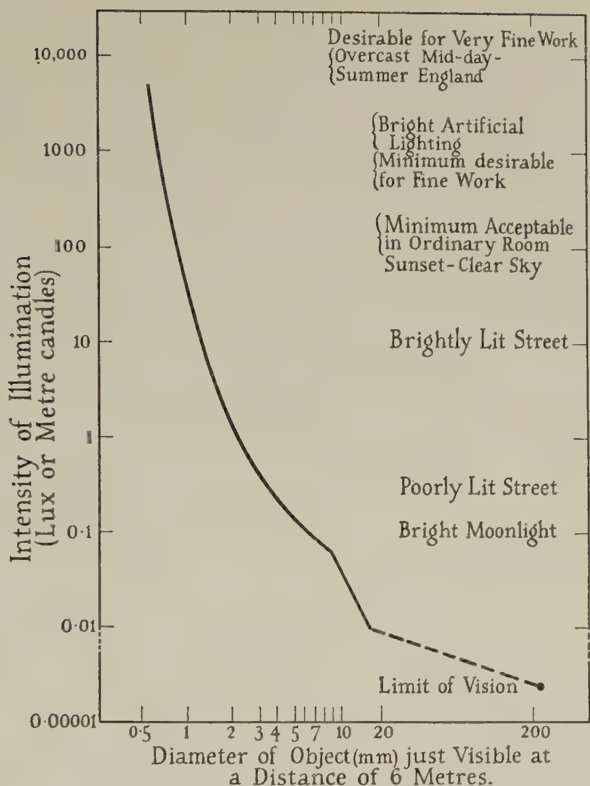


Fig. 18. 14. Variation of Acuity of Vision with Intensity of Illumination.

The diameter of a white object which is just visible at a distance of 6 metres is plotted against the intensity of the illumination (both on logarithmic scales). 1 lux = 1 metre-candle = 0·1 foot-candle. The visibility of an irregular object is the same as that of a symmetrical object of the same area.

An object 1·75 mm. diameter at a distance of 6 metres subtends an angle of 1 minute at the eye.

The acuity of vision is about doubled by increasing the illumination from that of an ordinary artificially lit room, to that of bright daylight (100,000 lux or more), and is increased some 400-fold from the minimum at which vision is possible, to bright daylight. The break in the curve indicates, approximately, the intensity at which cone vision becomes dominant. (From data by K. J. W. Craik.)

One of the complications is that the images formed on the retina cannot be perfectly sharp, partly owing to the defects of the optical system, and partly owing to diffraction. If, for example, the field of view consists of a number of alternate black and white lines of about the same width, light will spread out of the images of the white lines into those of the black lines ; the difference in illumination may fall below the threshold value, even if the images of the lines are as wide as the receptor units, and the presence of the lines will not be apparent. As the light intensity decreases, and the pupil dilates, the spreading of the image due to diffraction becomes smaller, but that due to spherical aberration in the optical system becomes larger : the net result, however, is that the image becomes less perfect as the illumination is decreased. But

since the size of the receptor units increases in these conditions, it is rather improbable that the imperfection of the image is of much consequence.

When the intensity is not much greater than the absolute threshold of vision, the random scattering of the relatively small number of light quanta falling on the retina may determine the acuity of vision. In the limit, for example, what will be seen will not be the pattern of light and dark regions which is formed on the retina by the cornea and lens, but the pattern formed by the way in which the quanta happen to be distributed on the retina.

Colour Vision. It is possible to recognise 120 different hues in the visible spectrum (400–800 mμ) and about 1,000 different shades altogether if one includes the purples, magentas and other colours which do not occur in the spectrum. Since 120 different hues can be distinguished in a range of 400 mμ, it follows that the eye can differentiate between wave-lengths that are less than 4 mμ different. The problem of colour vision is : how is this accomplished ?

It is, of course, conceivable that there might be 120 different kinds of retinal receptor, each connected to its appropriate class of nerve fibre, and associated with the sensation of a particular hue in the central nervous system. But this, on the face of it, seems unlikely, and the experimental observations on colour mixing provide evidence against it. If two beams of light are shone simultaneously into the eye—or if one looks at a white screen illuminated by the two beams—one of which is a pure spectral red in colour and the other a pure spectral green, it is found that by varying the relative intensities of these beams all the intermediate spectral hues of orange and yellow can be matched exactly ; and similarly mixtures of blue and green can reproduce all the spectral blue-greens. Further, it is found that by mixing red, green and blue in appropriate proportions, all known colours (including white) can be matched. These three colours are, therefore, called the *primary colours*. Two colours, such as a blue-green and a red, which can be mixed to give white are known as *complementary colours*.

It is to be noted that when we speak of colour mixing in physiology, we always mean mixing lights of different wave-length. If we coloured a white screen by means of a blue pigment, say, and illuminated it with white light, it would look blue because red light is hardly reflected at all, yellow light only slightly, and green light less than blue light. Similarly, a screen covered with a yellow pigment fails to reflect blue and violet light, and reflects red and green less than yellow light. A mixture of blue and yellow *paints*, therefore, fails to reflect either red or blue and violet, and reflects green more than yellow ; it looks green. For similar reasons, a mixture of red and green *paints* produces a dirty brown colour. But mixtures of suitable blue and yellow, or red and green, *lights* produce white. The mixture of two *subtractive* colours, produced by pigments illuminated by white light, produces quite different results from the mixture of two *additive* colours produced by illuminating a white screen with coloured lights.

The astonishing facts that two different stimuli such as red and green light can, when added to one another, produce a sensation which is qualitatively distinct from that produced by either stimulus alone, and that this sensation is identical with that produced by a yellow light, strongly suggest that the detection of colour is performed by means of

analysis into certain fundamental components. The observations on colour mixing indicate that no more than three fundamental components are needed. It would seem most probable that this analysis takes place in the retina, so that the actual optic nerve impulses are the same whether they are initiated by the yellow light or by the appropriate mixture of red and green lights. This is the basis of the Young-Helmholtz Theory of colour vision, first put forward by Thomas Young in 1801, and it still provides the most satisfactory explanation of the phenomena. This theory suggests that there are three colour sensitive systems in the retina corresponding to the three primary colours. These might be three photo-sensitive substances, perhaps contained in three different types of cone. As shown in Fig. 18. 15, one of these

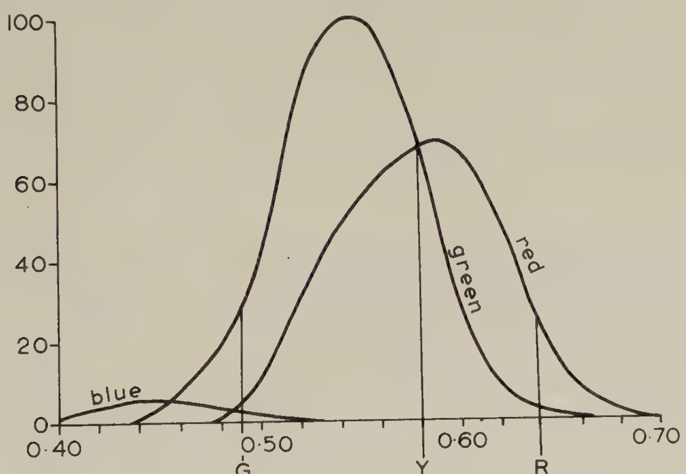


FIG. 18. 15. König and Dieterici's suggested **Response Curves** for three retinal receptors for colour vision (corrected by Judd).

Abscissæ—wave-length in μ.
Ordinates—relative sensitivities of three hypothetical receptor systems at different wave-lengths, the maximum sensitivity of the " green " system being made equal to 100. (From Walters : *Proc. Roy. Soc.*)

systems would respond maximally to blue light, less to green and not at all to red ; another would respond maximally to green light, less to blue and yellow and not at all to red ; and the third would respond maximally to red light, less to green and not at all to blue. A yellow light would stimulate both the " red " and the " green " systems in a ratio which depended on its wave-length ; that marked Y in Fig. 18. 15, for example, would stimulate them equally. One could also choose red and green lights and adjust their intensities so that they stimulated the " red " and the " green " systems, respectively, in any desired ratio ; those marked R and G on Fig. 18. 15 would stimulate them equally if they were of equal intensity. The combined stimuli from the red and green lights could thus be made to produce the same effect on the optic nerve as the stimulus produced by the yellow light, and the

resulting sensations would be the same. By extending this idea, we can state that any colour will stimulate the three retinal systems in varying proportions and that this stimulus can be reproduced exactly by a suitable mixture of three primary colours, red, green and blue.

Observations on colour mixing cannot, by themselves, provide a unique set of response curves for the three colour sensitive systems : there is a wide choice in the precise wave-lengths of the three pure colours which, when mixed in suitable proportions, will match a given colour. Additional evidence as to the nature of the response curves is obtained from observations on the changes in the apparent colour of a given source of light when the eye is made to adapt to lights of various colour ; on the anomalous results of colour mixing by people whose colour vision is defective in various ways (to be discussed in the next section) ; and on the changes in hue discrimination—*i.e.* the least change in wave-length which can be detected as a change in colour—in different parts of the spectrum. It is reasonable to presume that hue discrimination is best in those regions where the relative magnitudes of the responses of the three receptor systems is changing rapidly, *i.e.* where the curves in Fig. 18. 15 have the greatest difference in slope, as at about 600 mμ ; and is worst where they run most nearly parallel, as at about 540 mμ. The relative heights of the three curves must finally be so adjusted that the sum of all three has the same shape as the photopic luminosity curve (Fig. 18. 11).

These results from colour mixing experiments apply only when the field of view is relatively large, subtending 2 degrees or so : if very small fields are used, subtending 30 minutes or less, different results are obtained. Colour vision is less perfect, and the properties of the receptor systems vary with the size of the field and its position on the retina. If the central part of the fovea is illuminated with a small spot of light, it is found that the discrimination of hue is poor in the blue-green region of the spectrum, and the sensitivity to blue light, as compared with green or red light, is less than it is when large fields are used. All the colours of the spectrum can be matched by the use of *two* primary colours only ; colour vision in the central fovea is dichromatic, and not trichromatic as with larger fields, and is like that of the whole retina of certain people with defective colour vision, to be discussed below. The peripheral regions of the retina, on the other hand, are relatively more sensitive to blue light ; hue discrimination, although worse than that at the fovea in all parts of the spectrum, is best in the blue-green region. The observations with very small fields of view suggest that the receptor systems for colour vision interact over a relatively large area—of the order of 1 degree at least. Full colour vision is only obtained as a result of interaction between many receptor units, some of which must be outside the central fovea.

Defective Colour Vision. On the Young-Helmholtz theory it would be possible that one or more of the colour receptor systems might be absent or ill-developed. This does seem to occur in some people whose colour vision is defective ; but in others it would appear that all the systems are present and in action, but that their responses are mixed before reaching the brain, and so cannot be distinguished.

There are rare conditions of total colour blindness in which the whole

world is seen in shades of grey ; vision is *monochromatic*. In one of these the subject behaves at all intensities of illumination as normal people do at night, the brightest part of the spectrum being in the green ; he has a scotopic luminosity curve. In addition, the fovea is blind. These subjects, on attempting to fixate an object, move the eyes so that the image passes over the retina towards the fovea ; but since nothing is seen in this region, the eye overshoots the mark, moves in the opposite direction, overshoots the mark again, and so on. Rapid involuntary oscillations of the eye, of this kind, are called *nystagmus*, and this type of colour blindness is one of its many causes. In these subjects, the cones appear to be absent, or incapable of function, and they are accordingly known as *rod monochromats*. In the other type of monochromat, the subjects have photopic luminosity curves, not seriously different from those of normal people. It would appear that the receptor systems for colour are present, but that their responses cannot be sorted out. They are known as *cone monochromats*, and apart from the inability to distinguish colours, have no abnormal symptoms.

Slightly defective colour vision is not uncommon ; although commonly called colour blindness, the abnormality is often not detected until the colour vision is tested, and the subjects may go through life without being aware of the defect. These colour-defectives tend to confuse certain colours, most commonly reds and greens, a fact which has led to some disastrous railway accidents. The characteristic feature of the condition when fully developed is that all colours can be matched by mixtures of two spectral primaries, and vision is *dichromatic*, and not trichromatic : they are unable to distinguish differences in colour over large parts of the spectrum which look very different to normal people. The greater proportion of those with defective colour vision, however, are less seriously affected. They are trichromats, in that they need three primaries for colour mixture, but they are *anomalous*, since they have substantially less ability to detect differences in hue than have normal people. In colour mixing experiments, they do not accept a match made by a person with normal colour vision.

In the commonest types of dichromatic vision, known as *protanopia* and *deuteranopia*, the subjects are unable to detect any differences in hue in the red-yellow-green region of the spectrum. In protanopia, the photopic luminosity curve falls to abnormally low values at the red end of the spectrum, and it may be supposed that the "red" receptors are absent. But in deuteranopia, the photopic luminosity curve is not seriously different from that of people with normal vision ; it must be supposed that the responses from the "red" and "green" receptors are combined, owing to some permanent interconnections. The third type of defect, known as *tritanopia*, is rare ; hues cannot be distinguished in the blue-green-violet region. The luminosity curve is not seriously abnormal, and it appears that the responses from the "blue" and "green" receptors are combined. There are similarly three types of anomalous trichromats, and in the vast majority the deficiency is in the red-yellow-green region of the spectrum.

The detection of deficiencies in colour vision is not a task which can be undertaken light-heartedly, since there are all grades of severity, and when a defect is present in a mild form, it may only be possible to detect it when the conditions are made difficult. A subject may, for instance, show no signs until the area which he is required to recognise is very small, feebly illuminated, or the colour mixed with a lot of white light, as happens in a fog. In the Board of Trade test the candidate is required to name the colour of a small illuminated area of variable size. This is the basis of the Edridge-Green colour perception lantern. Another method is to present the suspect with a skein of wool, and ask him to pick out the one which matches it from a pile of variously coloured skeins. In the third method, which is the most convenient for rapid use, a series of cards (" Ishihara charts ") are printed with figures in coloured spots on a background of similarly shaped spots, but in a colour which is liable to look the same as that of the figure to a colour defective. The background spots may also be arranged to include a figure which is not obvious to a normal person, but which may be seen clearly by a colour defective.

Contrast Phenomena. When a black and white diagram is projected on a white screen, larger than the whole diagram, and the room has some background illumination, the black lines in the diagram appear darker than the surrounding parts of the screen. This is known as *simultaneous contrast.* The lines cannot, in fact, receive less illumination than any other part of the screen (the projecting lantern cannot emit black light) : and we must suppose that the signals from the receptors in the retina which are under the images of the black lines are suppressed, or inhibited, by the signals from the nearby receptors which are strongly illuminated. Electrophysiological evidence shows that such mutual inhibition of one receptor cell, or group of cells, by another does actually occur (p. 559 and Fig. 18. 18). This effect is almost certainly one of the complicating factors concerned in limiting visual acuity, referred to above (p. 542). An inhibited receptor unit will give no response to any light that may be shining on it : even though the dark gap is not completely black, it will appear so, and the minimum detectable width will be reduced.

The effect of simultaneous contrast is well seen, also, if a small piece of grey paper is placed on a large square of red paper ; the grey appears tinged with complementary green. Each colour can evoke the complementary colour sensation. Strongly illuminated receptors suppress signals from neighbouring weakly illuminated receptors only if they are of the same kind (have the same colour sensitivity) as themselves.

If one looks steadily at a small piece of coloured paper for a few seconds, and then at a large piece of grey paper, the space previously occupied by the coloured paper is now seen in the complementary colour. This is the commonest type of *after-image*, and the phenomenon is known as *successive contrast* : it suggests that each of the receptor systems for colour may become " fatigued " or " adapted " independently of the others. After-images, however, may also be of the same

colour, more or less, as the stimulating light : it must be supposed that the " recovery " of the stimulated receptor system is delayed in some way. But little is known about the origin of after-images ; there is reason to suppose, however, that they result in part, at least, from changes that occur in the retina.

Flicker and Persistence of Vision. When light of a certain intensity is suddenly shone into the eye, its apparent brightness at first increases with time, as might be expected from the existence of temporal summation. But if the intensity is well above threshold, the apparent brightness rises to a peak and then declines, reaching a steady value in a few tenths of a second to a second : the greater the intensity, the higher is the peak in relation to the steady value, the sooner is it reached, and the more rapidly does the brightness decline. This indicates that many of the receptor units, particularly when the intensity is high, must be rapidly adapting, in the sense defined in the previous chapter (p. 508), and do not respond to maintained stimulation. (This rapid type of adaptation must not be confused with the much slower kind, peculiar to the eye, and described as light- and dark-adaptation.) The additional sensitivity when the intensity is changing rapidly is partly responsible for the fact that flashing or moving lights are more obvious, and attract more attention, than steady ones. This effect will be referred to again later.

These measurements are made by presenting the eye with a field of view which is divided into two parts : one of these is illuminated steadily, and the other with flashes of known duration. By varying the relative intensities of the two parts, so as to keep them matched, it is possible to discover how the apparent brightness varies with the duration of the flash, and thus with time after the beginning of the illumination.

The sensation produced by a short flash of light outlasts the stimulus, so that if a series of flashes separated by short intervals of darkness is presented to the eye, the resultant sensation may be one of continuous illumination. This is known as the " persistence of vision." The frequency at which the flashes must be made to follow one another in order that the illumination should just appear continuous, and not flickering, is known as the " critical frequency of flicker." The brighter is the flash, the more quickly does the sensation die away ; if the frequency has been adjusted to the critical value for one intensity of illumination, the flicker will reappear if the intensity is increased. This method can thus be used to measure the subjective brightness of lights of different intensity : it is useful for measuring the relative subjective brightness of different parts of the spectrum, where the judgments are complicated by the presence of colour.

In cinematograph projection there is a dark interval on the screen whilst one " frame " is being changed to the next, but owing to the persistence of vision these dark intervals are not appreciated. In most conditions the peripheral parts of the retina are more sensitive to flicker than the central parts. This can be demonstrated in suitable circumstances by looking away from the screen, when the flicker will be observed.

The exact relation between the critical flicker frequency and the intensity

of the light varies with the intensity, with the area and region of the retina illuminated, and with the state of dark- or light-adaptation of the eyes. Care is necessary, therefore, in making quantitative use of the relation.

Photochemical Reactions in the Retina

It is known that light, as such, is incapable of stimulating a nerve. It can do so only through some receptor which, according to our present ideas, must contain a *photochemical substance*—that is, a substance whose chemical composition, or internal molecular or electronic structure, is changed by the action of light. This change, moreover, cannot occur unless the light is absorbed, and the amount of change produced is likely to depend on the quantity of light absorbed. The sensitivity of the eye, both in man and in all animals investigated, varies with the wave-length of the light incident on it, as is illustrated in Fig. 18. 11. It is to be expected, therefore, that the fraction of the incident light which is absorbed by the photochemical substance (its optical density) will also vary with the wave-length ; it will thus appear coloured. Such coloured substances are present in the retina, and they are clearly affected by light since the colour (and optical density at appropriate wave-lengths) is greatly diminished when the retina is illuminated. Substances concerned in photochemical reactions do not necessarily undergo any change in the optical properties : the fact that those concerned in vision are bleached by light has been of great assistance in detecting their presence and in studying their properties.

Scotopic (Rod) Vision. If the retina of a dark-adapted animal is removed, it is found to be deep pink in colour, and this colour changes rapidly on exposure to light, so that after a few minutes, an isolated retina becomes yellow or white. If the dark-adapted retina is extracted by solutions of digitonin, bile salts or certain other hæmolytic agents (*i.e.* those which break up cells), a deep pink solution is obtained which is bleached by light in the same way as is the retina itself. The photosensitive pigment contained in the solution is known as *visual purple*, or *rhodopsin*. Microscopic examination, either of the fresh retina of a dark-adapted frog, or of a suitably fixed and stained mammalian retina, shows that rhodopsin is contained only in the rods.

Rhodopsin consists of a reacting, or chromophore, group attached to a protein. The action of light is, at first, to bring about various intramolecular changes affecting the bond between the chromophore and the protein, so that substances with different absorption spectra are formed : most of these have only a very transitory existence, and are orange or yellow in colour. Eventually, nearly all colour is lost, and the bond between the chromophore and the protein is loosened, so that the chromophore becomes more readily soluble in organic solvents. The substance extracted by such solvents is called *retinene*, and has been shown to be vitamin A aldehyde, an oxidation product of vitamin A, which is itself an alcohol. Rhodopsin, therefore, is a compound of retinene with a protein to which the name *opsin* has been given : it is a phospho-protein and contains phospholipin.

The photochemical reaction by which rhodopsin is bleached is reversed, spontaneously, to a small extent when the light is removed ; but the regeneration of rhodopsin is always very incomplete, only a minute fraction of the original quantity being regained. This is because the retinene which is formed by the action of light on rhodopsin has a different molecular configuration from the retinene which is formed by the oxidation of vitamin A, and will not combine with the protein carrier, opsin. In surviving mammalian eyes, after bleaching by light, there is little or no regeneration of rhodopsin unless the blood supply is intact : there is some regeneration in excised eyes of cold-blooded animals, but only in the presence of oxygen, and provided that the retina has not been detached from the pigment cell layer which ordinarily surrounds it. Further, it has been shown that animals which are deprived of vitamin A either do not regenerate rhodopsin at all, or do so very slowly. There is, also, a form of night blindness, as already mentioned (p. 536), which occurs among people living under conditions of extreme privation, and which can be cured by adding vitamin A to the diet. It is clear, therefore, that for the re-synthesis of rhodopsin, after it has been bleached, a fresh supply of the correct isomer of retinene is needed. This, for the most part, is obtained from vitamin A, which is oxidised in the course of normal metabolic reactions, the phosphopyridine nucleotide coenzyme (DPN) being essential as an intermediary hydrogen acceptor : but there may, also, be some direct conversion of one isomer of retinene to another, the reaction being catalysed by appropriate enzymes.

The fact that rhodopsin is pink in colour means that it must transmit red and blue lights and absorb green light : it follows that the wavelengths which appear green must be more potent in bleaching rhodopsin than those at the ends of the spectrum. The more precise relation between the optical density of rhodopsin and the wave-length of the incident light (its *absorption spectrum*) and the relation between the rate of bleaching and the wave-length of the light used (its *action spectrum*) have both been accurately measured ; they are identical. If a light of known wave-length is made to shine into the human eye, it is possible by the use of suitable refined and delicate methods to measure the fraction of the incident light which is reflected from the back of the eye after passing twice through the retina. The measurements show that in the extra-foveal region of the retina of a dark-adapted man, there is a pigment which is bleached when the eye is strongly illuminated. This pigment appears to be contained in the rods, since its optical density in different regions of the retina varies with the number of rods per unit area, as observed histologically. Subtraction of the absorption spectrum of the bleached retina from that of the unbleached retina gives the *difference spectrum* of the photo-sensitive substance : this is shown in Fig. 18. 16, and resembles very closely the difference spectrum of rhodopsin. (This is not quite the same as its absorption spectrum, since bleached rhodopsin absorbs violet light rather more than light of other colours.) The retinal pigment, moreover, is not

bleached by red light and, for a given intensity, it is bleached most rapidly by green light : its action spectrum is thus similar, at least, to that of rhodopsin. Further, the difference spectrum of the retinal pigment agrees very closely with the curve of the sensitivity of the dark-adapted (scotopic) eye to the different parts of the spectrum, as may be seen by comparing Fig. 18. 16 with Fig. 18. 11. Thus, both rhodopsin and the retinal pigment absorb green light more completely than red light ; and are rapidly bleached by green light but unaffected by red light. In rod vision, the sensitivity is greatest to wave-lengths which appear green if the colour can be appreciated, and wave-lengths

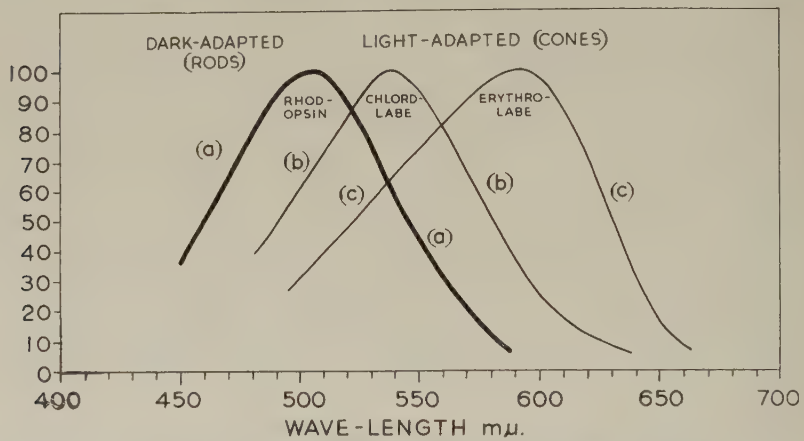


FIG. 18. 16. Photo-sensitive Pigments in the Human Retina.

Three difference spectra are shown, as obtained by bleaching the retina :

(*a*) in the dark-adapted extra-foveal region, indicating the presence *of rhodopsin* in the rods;

(*b*) and (*c*) in the light-adapted fovea, indicating the presence of two other pigments, *chlorolabe* and *erythrolabe* in the cones.

The peaks of the curves are at 505 mμ for rhodopsin ; 540 mμ for chlorolabe ; and 590 mμ for erythrolabe. The ordinates have been arbitrarily adjusted so that the peaks are brought to 100 in each case. (From Rushton.)

which would appear red cannot be seen at all. Or, to put it in another way, suppose that we have a source of light whose wave-length can be varied, and that we adjust its intensity so that, whatever the wave-length, it has the same brightness as observed by a dark-adapted man (no colour, of course, will be appreciated) : then, in a given time, whatever the wave-length used, it will bleach the same amount of rhodopsin and the same amount of the pigment in the retina. There is thus very strong evidence that rhodopsin is the photo-sensitive substance which is responsible for human scotopic vision.

The peaks of the retinal difference spectrum (Fig. 18. 16) and of the scotopic luminosity curve (Fig. 18. 11) are at about 505 mμ ; that of the absorption spectrum of rhodopsin is very close to 500 mμ. These small discrepancies have two different origins. The shift towards the red of the retinal difference

spectrum is due to the necessity of bleaching the retina rapidly by a strong light ; an appreciable amount of a transient intermediate substance with an orange colour is still present when the measurements are made. The shift of the luminosity curve is produced by the faint yellow colour of the human lens, so that the whole eye is slightly less sensitive to blue light than is the retina itself. If the appropriate corrections are made, the retinal difference spectrum and the scotopic luminosity curve agree very closely with the difference spectrum of human rhodopsin.

Photopic (Cone) Vision. It has never been possible to observe in the isolated retinæ of the frog or of any mammal, or in extracts made from them, the presence of pigments which might act as photo-sensitive substances for vision at relatively high intensities of illumination. Study of the human fovea, however, by an appropriate modification of the method used to detect the presence of rhodopsin, shows that photo-sensitive pigments are present.

If a strong white light is used for bleaching the foveal region of the retina, the difference spectrum resembles, in shape and position on the wave-length scale, the photopic luminosity curve, given in Fig. 18. 11. If a deep red light is used, the difference spectrum is now found to correspond more with the response curve of the hypothetical " red " receptor system, as given in Fig. 18. 15. This suggests that the retina contains a red-sensitive pigment, which has been called *erythrolabe* (" red-taking "). Subtraction of this curve from that obtained by bleaching with white light leaves a curve which is similar in shape and position to the response curve of the hypothetical " green " receptor system. The green-sensitive pigment has been called *chlorolabe* (" green-taking "). The difference spectra of these two pigments are shown in Fig. 18. 16. More direct evidence for the existence of a green-sensitive pigment is derived from observations on subjects with the kind of defective colour vision known as protanopia. The difference spectrum of the pigment in the fovea of such a subject is independent of the colour of the light used for bleaching, and is sensibly identical with that of chlorolabe ; the red-sensitive pigment, erythrolabe, is absent. This curve may be superimposed, very closely, on the photopic luminosity curve of the protanopic subject. As remarked above (p. 547), the shape of the luminosity curve, and the nature of the anomalies of colour vision, suggest that in protanopia the " red " receptor system is missing. There is, as yet, no indication of a pigment corresponding to the " blue " receptor system (*cyanolabe* or " blue-taking ") : the relatively small area of the response curve, however, would be expected to increase considerably the technical difficulties of observing it. On the whole, therefore, there is good evidence for the existence of photo-sensitive substances in the human retina whose properties correspond closely with those to be expected on the basis of the Young-Helmholtz theory of colour vision.

The photopic luminosity, or magnitude of the visual sensation at any wave-length, cannot necessarily be compared directly with the difference, or action, spectra of the three photochemical pigments concerned. It will vary with the wave-length of the light in the same way

as the combined optical densities of the pigments only if the "efficiencies" of the three systems are identical—that is, the amount of light which must be absorbed in order to produce a sensation of a given size must be the same in all. There is no necessity that this should be so; and measurements indicate that the chlorolabe system is substantially more "efficient" in this sense than is the erythrolabe system. The rhodopsin system in the rods is more "efficient" than either; but since only one photochemical pigment is involved, the "efficiency" does not affect the comparison between the shape and position of the scotopic luminosity curve and the difference spectrum of rhodopsin.

A great many different photo-sensitive pigments have been observed in the eyes of animals other than man. Some of these have been shown to consist of a combination of retinene with opsin, and it seems likely that most, or perhaps all, are essentially similar to rhodopsin in this respect. There are, however, two different kinds of retinene, corresponding to the two different kinds of vitamin A, with slightly different chemical structures, and each of these has several isomers with different internal molecular configurations, only two of which will combine with opsin. There are, also, at least two (and probably many more) different kinds of opsin. The absorption spectra of all these pigments have the same shape, as far as is known, but the position on the wave-length scale depends on the nature of the retinene and on the nature of the opsin with which it is combined. It is probable, indeed, that even in man, no two individuals have quite the same opsin; there are certainly appreciable individual differences in the shapes and positions of the luminosity curves.

Many different kinds of animal, other than mammals, appear to be able to distinguish colours, but the methods used may be different. If the retina of a frog is examined under the microscope, two different kinds of rod can be seen: the most numerous are pink in colour and contain rhodopsin, but in addition, there is a small proportion of rods which are green in colour and contain a photo-sensitive pigment which has a maximum absorption in blue light. The retinæ of domestic fowls, like those of diurnal birds in general, contain mostly cones, with relatively few rods. A photo-sensitive pigment has been extracted, called *iodopsin*, which is a retinene-opsin compound, and whose absorption spectrum has much the same shape and position on the wave-length scale as has the human photopic luminosity curve. The cones of birds contain coloured oil droplets which may act as light-filters, and differentiate the cones into two or more classes with different colour sensitivities.

Adaptation and the Photochemical Steady State. In the living eye, the photo-sensitive pigments which are bleached by light must obviously be continuously regenerated, otherwise vision would very soon cease. The rate of bleaching is proportional to the intensity of the light, as already remarked; while the rate of regeneration increases with increase in the fraction of the pigment which is in the bleached condition. This, indeed, is to be expected, since the opsin at least is used again, and the rate of its combination with retinene will depend on its concentration. In any conditions of steady illumination, therefore, the fraction of the pigment bleached will continue to increase until the rate of regeneration has become equal to the rate of bleaching. There will then be a *photochemical steady state*, in which the fraction of the photo-sensitive pigment which is in the bleached condition will remain constant, at a value which increases with increase in the intensity of illumination.

Measurements of the optical density of the pigments in the human retina show that the rate at which the photochemical reaction approaches the steady state, both in the rods and in the cones, is very similar to the rate at which the sensitivity of the eye changes during dark- and light-adaptation. If the intensity of illumination is large, most of the photo-sensitive pigment will be in the bleached condition, the optical density of the remainder will be small, only a small fraction of the incident light will be absorbed, and the sensitivity of the eye will be low. If, now, the subject goes into the dark, the pigment will be regenerated, a greater fraction of the incident light will be absorbed by it, and the sensitivity will rise. This was at one time considered to be the chief source of the changes in sensitivity during dark- and light-adaptation. But it is now known that quantitatively the changes in the density of the pigment are far too small to account, by themselves, for the changes in the sensitivity of the eye. If, for example, the pigments are almost entirely bleached away by illumination with a strong light, and the eye is then left in the dark, nearly all the pigment in the cones is regenerated in about seven minutes, and nearly all the rhodopsin in the rods is regenerated in about thirty minutes : these times correspond reasonably well with the times taken for dark-adaptation to occur, as shown in Fig. 18. 12. But if the density of rhodopsin is measured after the eye has been in the dark for seven minutes, it is found to be just one-half of the maximum density reached after thirty minutes or more. One would expect, therefore, that the visual threshold would then be twice the absolute threshold reached after the eye has been in the dark for so long that no further change occurs. In fact, after about seven minutes in the dark, the visual threshold is at least 500 times the absolute threshold : the magnitude of the change in sensitivity during the remainder of the dark-adaptation process is thus quite out of proportion to the magnitude of the change in optical density of rhodopsin. Some such discrepancy, of course, is to be expected owing to the change in the size of the receptive field, as discussed on p. 539, above. But this effect may be eliminated by measuring the threshold when only a very small area of retina is illuminated, subtending only a few minutes of arc : even then the threshold, after seven minutes in the dark, is some 50 times the absolute threshold, instead of twice, as expected. The increase in the retinal action time during dark-adaptation will also contribute to the reduction in visual threshold ; but, as may be seen from Fig. 18. 13, the increase is far too small to account for the observed discrepancy.

The sensation of vision, then, is not simply and directly related to the quantity of light absorbed in the photo-sensitive pigment, and the amount of the pigment bleached during the retinal action time. We must suppose that, as the light intensity increases, and the eye becomes light-adapted, the receptor cells become progressively less sensitive. In the fully dark-adapted state a rod is activated by the absorption of a single quantum of light energy. Each rod is estimated to contain between 10 million and 1,000 million molecules of rhodopsin, but only one of these can be directly affected by the single quantum. It would

seem likely that some amplifying or triggering process must be set into action by the bleached rhodopsin molecule. When the eye is partially or fully light-adapted, a great many quanta must be absorbed before the rod, or cone, becomes activated ; the sensitivity of the triggering process may perhaps vary with the state of adaptation of the eye, and change considerably with small changes in the proportion of the rhodopsin in the bleached, and unresponsive, condition. Studies with the electron microscope, and with polarised light, show that the rods are highly organised internally, with the rhodopsin molecules arranged in an orderly way in a number of discrete compartments. It may be that if any one rhodopsin molecule absorbs a quantum of light, it affects all the others in the same compartment : if there are enough unbleached molecules in this compartment, the rod becomes activated ; if not, several compartments must each absorb a quantum. But at present this is little more than speculation.

Electrical Phenomena and Nerve Impulses

The Electroretinogram. If an electrode is placed on the cornea, and another on the back of the eye opposite to it (assuming the eye to have been removed), potential differences are developed between them when light is shone on the eye. If the eye is *in situ* as in an experimental animal or human subject, the second electrode may be placed at any convenient point on the skin, usually on the forehead in man. The potential changes, as recorded on a moving surface, are of complex form, and are known as *electroretinograms*. Typical records from experimental animals are shown in Fig. 18. 17 : an upward deflection indicates that the electrode on the cornea, or in effect the nervous layer of the retina, becomes electrically positive to the indifferent electrode, or in effect the rod and cone layer.

In a general way, the size of the electroretinogram increases as the intensity, area and duration of the illumination are increased ; the last two only within limits similar to those which apply to the sensation of brightness in man. Its general shape is very much the same whether the primary receptor cells that are in action are rods or cones : but in detail, the shape may change according to the state of dark- or light-adaptation of the eye, and according to the kind of animal from which it is obtained. If the intensity and duration of illumination are made sufficiently large, the main initial positive peak (*b*-wave) reaches a limiting size, and on further increase may become smaller ; the slow positive wave (*c*-wave) becomes less evident and may vanish ; and the deflection which occurs just after the illumination ceases (the " off-effect " or *d*-wave) becomes relatively larger. The " off-effect " is more prominent in light-adapted than in dark-adapted eyes, and in eyes which contain many cones : it is well seen in the frog, but not in the cat, for example.

Human electroretinograms are very similar, on the whole, to those of other vertebrates. The slow maintained potential (*c*-wave) and the " off-effect " (*d*-wave) become obvious only when a rather long period of

light stimulation is used, and in man, it is difficult to avoid movements of the whole eye, of the pupils and of the eyelids (blinking), all of which may introduce electrical artifacts. For this reason, brief flashes are often used ; the initial short negative *a*-wave and the subsequent positive *b*-wave are then recorded, but the *d*-wave (if present) may be confused with the *b*-wave.

The size of the *b*-wave varies with the wave-length of the light used, other things being equal. If the eye is dark-adapted and the intensity of illumination small, the curve relating the size of the *b*-wave to the wave-length of the light (the intensity being constant) is very similar

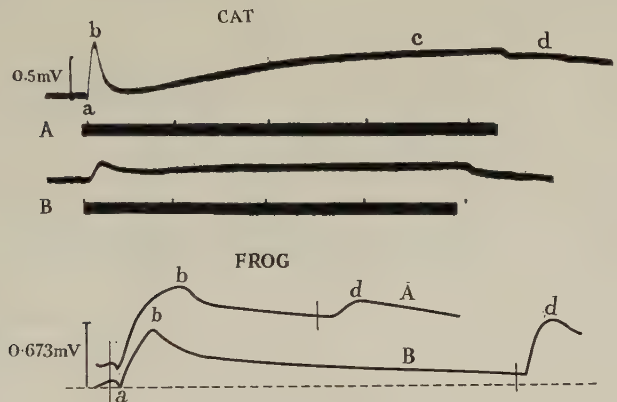


Fig. 18. 17. Electroretinograms of the Cat and the Frog.

Upper two records from the cat, showing the effect of increasing the intensity of illumination, the intensity in record A being 100 times that in record B. The dark lines below the records indicate the periods of illumination, with time marks at 0·5 second intervals. (After Granit.)

Lower two records from the frog, showing the effects of dark and light adaptation. The short vertical lines show the beginning and end of illumination.

Record A—dark adapted, illuminated for one second.
Record B—light adapted, illuminated for two seconds.
(After Granit and Riddell.)

to the scotopic luminosity curve in man, and to the spectral absorption and action curves of rhodopsin. If the eye is light-adapted and the intensity large, the curve follows the photopic luminosity curve. (This can only be demonstrated in eyes in which there are many cones : the *b*-wave becomes too small in light-adapted eyes of cats, for example.)

Attempts have been made to find the origins of the various components of the electroretinogram by inserting micro-electrodes into the retina to various known distances from the inner surface. Electrical changes are observed which are similar in nature to those recorded by the more conventional large electrodes, but varying according to the position of the electrode. The structure responsible for each of the component waves, however, has not yet been conclusively identified : the pigment epithelium, the receptor layer and the inner nuclear layer

(bipolar cells) all seem to generate electrical potentials, but none of the waves can be definitely regarded as a receptor potential, as described in the previous chapter.

Action Potentials. The overall activity of the whole optic nerve can be observed by means of electrodes placed on it, in suitable kinds of experimental animal. As one would expect, the relation between the intensity of the stimulus (in this case light) and the number of impulses passing up the optic nerve in a given time is similar to that for other receptor systems. Within limits, the frequency of the optic nerve impulses appears to depend on the total *amount* of light stimulating the eye. In other words, in order to obtain a given response the product of the intensity of the light, the duration of the illumination and the area of retina illuminated, must be kept constant. This, as we have seen, applies also to the electroretinogram and to the subjective sensations. The frequency of the impulses, moreover, like the subjective sensation of brightness as described on p. 549, above, rises with time after the start of the illumination, reaches a peak and then declines : the greater the intensity, the higher the peak, the sooner it is reached and the more rapidly it decays. This is a clear indication that a great part of the rapid adaptation must occur in the retina and not in the central nervous system.

The subjective sensation of brightness (as observed in man) thus seems to follow in a general way the frequency of the impulses in the whole optic nerve (as observed in experimental animals). But the more detailed content of the sensation—the detection of spatial and temporal patterns, and of colour—must be conveyed by signals in the separate nerve fibres from different receptor units.

By the use of micro-dissection methods, and of micro-electrodes, it is possible to record the action potentials in single optic nerve fibres, the ganglion cells from which they arise, and (probably) the bipolar cells. These experiments have shown that in both the frog and the cat, there are two main types of response. The impulses may increase in frequency when the light is switched on, and decrease in frequency, or cease altogether, for a short time after the light is switched off (the " on " response, Fig. 18. 18, A) : or the impulses may decrease in frequency, or cease, when the light is switched on, and increase in frequency after it is switched off (the " off " response, Fig. 18. 18, B). This " off " response can be suppressed immediately by turning the light on again. The impulses usually appear at a high frequency for a fraction of a second after the light is switched on, or off, and then settle down to a steady lower frequency, or may cease. Some nerve fibres give a short burst of impulses both at " on " and at " off " (the " on-off " response). This type of response is characteristic of the rapidly adapting type of receptor ; the response is determined by the *change* in the illumination, whether this is an increase or a decrease, as for example, would accompany movements of light and shade across the retina. Such movements, particularly in the peripheral regions, are particularly effective in attracting attention.

Observations on single fibres have demonstrated conclusively a fact that may be inferred from other evidence, and has already been mentioned. The receptive field connected to a single nerve fibre may be very large, and action potentials may be set up in a given nerve fibre by illuminating rods or cones spread over a wide area ; the sensitivity decreases markedly towards the periphery of the field, so that it has no sharp boundary. Conversely, a very small spot of light, stimulating only a very few receptor cells, will produce impulses in several different fibres, so that the receptive fields of different units must overlap considerably.

If the eye is fully dark-adapted, there is summation between the effects of illuminating any one part of the receptive field with that of

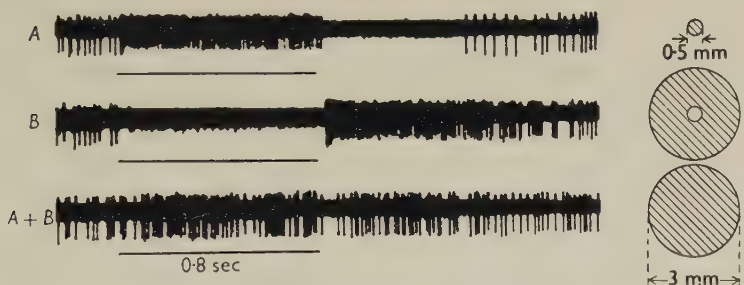


FIG. 18. 18. Action Potentials from a cell in the inner nuclear layer of a cat's retina.

A. On illuminating a spot 0·5 mm. diameter in the centre of the receptive field of the cell ; " on " response.
B. On illuminating a ring 3 mm. diameter surrounding the 0·5 mm. spot illuminating in A ; " off " response.
A + *B*. On illuminating both the central spot and the peripheral ring ; both responses are present but each is partially inhibited by the other.
The patterns of light falling on the retina are shown on the right of the figure. (Brown and Wiesel, 1959. *J. Physiol.*)

illuminating any other part : the threshold light intensity falls as the area illuminated rises, and Ricco's law (p. 539, above) is obeyed in the central part of the whole receptive field by some units, but not by all. If the eye is not fully dark-adapted, the discharge set up by illuminating one spot in the receptive field may be inhibited when another spot is illuminated simultaneously : the discharge obtained from a unit in the centre of the receptive field, when that part is illuminated, is regularly inhibited when the peripheral part is illuminated as well, and *vice versa*. This is illustrated in Fig. 18. 18. In the cat, one and the same unit may give, say, an " on " discharge when the central part of its receptive field is illuminated, and an " off " discharge when the peripheral part is illuminated (as in Fig. 18. 18), or *vice versa* : this does not affect the mutual inhibition.

The reduction in the size of the receptive field as the intensity of illumination is increased, as described above (p. 539), results, therefore,

not only from a failure of mutual summation between receptors in the central and peripheral parts of the field, but also from a development of mutual inhibition. The mutual interaction between receptor units seems to occur largely at the level of the bipolar cells ; there is the same complexity and variability in their responses as in those of the ganglion cells.

Spectral Sensitivities. The least amount of light energy which will produce a just detectable response in an optic nerve fibre varies with the wave-length of the light. If the eye examined is dark-adapted, the spectral sensitivity of all the receptor units has a maximum at about 500 mμ, the curve has much the same shape as the human scotopic visibility curve, and is just what would be expected if rhodopsin were the photochemical substance in action. If the eye is light-adapted, and if the retina contains a sufficient proportion of cones, the spectral sensitivity of some of the units will show a Purkinje shift ; the maximum sensitivity now being at about 560 mμ, and the curve resembling the human photopic visibility curve. Granit, who has been responsible for much of this work on single optic nerve fibres, called these curves the scotopic and photopic *dominator curves.* But in addition, light-adapted eyes are found to contain units whose spectral sensitivity curves have more complex shapes, some with humps and others with sensitivity maxima at two different wave-lengths ; some curves, also, even though they have only one peak, are much narrower than the dominator curves. These results are explained by Granit as being due to the existence of *modulator curves,* of which there are three kinds : those with maxima in the red part of the spectrum (about 600 mμ) ; those with maxima in the green (about 530 mμ) ; and those with maxima in the blue (about 450 mμ). The precise position of the maximum sensitivity varies somewhat from unit to unit, but always lies in one or other of these three regions. The photopic dominator curves may be presumed to be derived from units which add together, in suitable proportions, the responses from all the three kinds of receptor which are inferred to exist from colour-matching experiments, and from observations on the difference spectra of the human retina. Such units will be responsible for a sensation of general brightness. The modulator curves may be derived from units in which the response of one class of receptor inhibits that of another : the final response at any given wave-length will thus be the difference between the responses of the two different kinds of receptor. Such units may be responsible for the sensation of colour, although further analysis must occur in the central nervous system.

Eye Movements

The eye is supported in such a way that it is allowed very free movement in all directions. The eyeball, its muscles, nerves and blood vessels are lodged in the orbit, the rest of the orbital cavity being filled by a mass of fat. It is supported by the suspensory ligament, a thickening of the tough loose-fitting envelope (Tenon's capsule) which is

attached to the sclera just behind the corneo-scleral junction, and which sends prolongations to blend with the sheaths of the muscles and nerves. The whole eyeball is rotated about vertical and horizontal axes through its centre, to left and right, or up and down, by the action of the *extrinsic muscles*, as shown in Fig. 18. 19.

The *internal rectus*, on contraction, turns the eye inwards towards the nose, and the *external rectus* turns it outwards. The *superior rectus* and the *inferior rectus* turn the eye upwards and downwards, respectively. But since the optic foramen, where these muscles have their origin, is on the medial side of the orbit, they do not lie in the same vertical plane as the optic axis of the eye, and thus produce horizontal movements

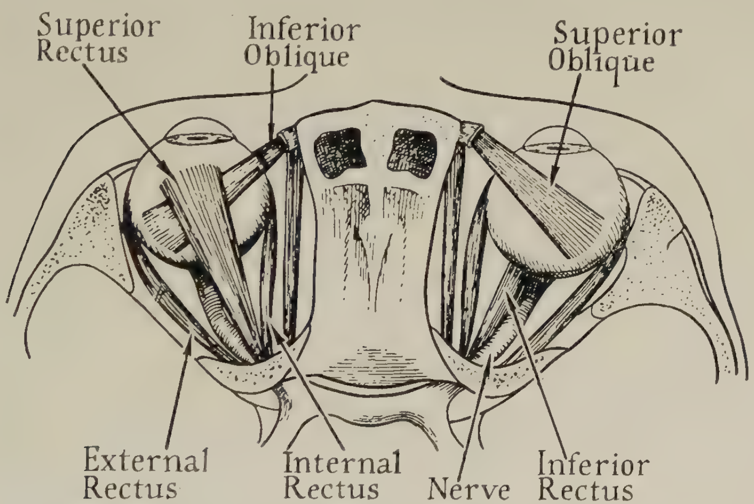


FIG. 18. 19. **Oculomotor Muscles** in Man as seen from above.

On the left, a portion of the superior oblique muscle has been cut away to reveal the inferior oblique ; on the right, the superior rectus muscle has been removed, so as to show the inferior rectus. (Modified from Adler.)

and rotations as well as vertical movements. These are counteracted by appropriate contractions of the *inferior* and *superior oblique* muscles. The superior rectus turns the eye inwards, as well as upwards, and rotates the right eye clockwise and the left eye anti-clockwise, when looked at from the front. The inferior oblique turns the eye upwards and outwards, and rotates the right eye anti-clockwise. When the superior rectus and the inferior oblique contract simultaneously, the eye is moved vertically upwards, the other movements cancelling one another out. The inferior rectus turns the eye downwards, but also inwards, and rotates the right eye anti-clockwise. The superior oblique passes through a bony hole, the trochlea, situated to the upper medial side of the front of the orbit : by pulling from this point, it turns the eye downwards and outwards and rotates the right eye clockwise. The sideways and rotational movements produced by the inferior rectus

are thus opposed, and acting together they produce a pure downwards movement of the eye.

Our attention is directed to some object in the external world most usually either as a result of some sound reaching our ears ; or as a result of its image appearing on some peripheral part of the retina, or, if already there, moving or changing in some way. Each eye is then moved so that the image of the object concerned falls on the fovea and is seen most distinctly. This may be a conscious act, but may also be reflex ; the continued and delicate adjustment of the excitation and inhibition of the extrinsic muscles, necessary to keep the image on the fovea, however, is certainly reflex. The control is initiated chiefly by means of the signals from the retina itself ; an object moving in the field of view, or a certain part of a whole moving field, for example, can thus be followed accurately. But an important and frequent source of movement of the image on the retina is movement of the head in space, consequent, usually, on movement of the whole body : this is detected by the labyrinthine organs (Chapter 17) and signals from these organs assist in bringing about the necessary compensatory movements of the eyes.

The adjustment of the direction in which the eye is looking is extremely precise, but is never perfect, even though there are no complicating movements either of the head or of the object looked at. Unless special attention is given to maintaining an exact fixation, the image on the retina is continually shifting slightly, by a few minutes of arc at intervals of a few tenths of a second. This has, in fact, some advantages. Acuity of vision depends on the detection of small differences in the illumination falling on neighbouring cones : but the electrophysiological evidence suggests that many of the receptor units behave as if they were rapidly adapting. Slight movements of the image from one to another would thus facilitate its detection. Colour vision, also, necessitates the illumination of relatively large numbers of receptor units, either simultaneously or in rapid sequence : the anomalous types of colour vision associated with very small areas of illumination can only be detected when fixation is very careful and deliberate.

Vision with Two Eyes. The field of vision for one eye is more than 180 degrees. With two eyes arranged on the sides of the head, an animal can therefore see all round itself. In man, the eyes look directly forwards, so that the total field of vision is little more than 180 degrees ; the advantage of seeing parts of the outside world simultaneously with the two eyes has entailed a loss in the total size of the field of vision.

The field of vision is mapped on an instrument known as a *perimeter*. An object is kept at a constant distance from the eye, and arranged to move round an arc, with the eye as centre, above or below the line of vision, or on the nasal or temporal side. The greatest angle at which it can be seen defines the limit of the field of vision. Mapping the field of vision in this way is of value in detecting blind areas due to accident or disease in the eye itself or in the nervous pathways for vision.

Certain portions of the outside world form almost identical images on the two retinæ, and both these images must be transmitted along the optic nerves. There is no doubt that we are conscious of both images simultaneously, because two sets of objects can be arranged each hidden to one eye but visible to the other, and in these circumstances all the objects are seen together. This faculty of blending the two visual images into one conscious picture is known as *binocular vision.* By suitable devices, entirely different pictures can be presented to each eye and normally we are conscious of both, but when the need arises it is possible to suppress one image, as for instance on looking down a microscope with both eyes open. There are people, however, who have not got this faculty of fusing the images from the two eyes, and habitually suppress one image. If an ordinary person prevents the almost identical images from falling on corresponding points of the retina, as by pressing on the eyeball, then he sees double (diplopia).

The maintenance of binocular vision demands very precise co-ordination between the movements of the two eyes. In particular, in order to look at a near object, the optic axes of the two eyes must be made to *converge.* The stimulus for this to occur is the presence of two similar retinal images. Convergence, as already mentioned (p. 530, above), is very closely connected with accommodation and constriction of the pupil, all three occurring together when a near object is viewed, and being ordinarily parts of one and the same reflex.

Convergence and accommodation can be dissociated. If a weak prism is placed in front of one eye so as to cause an apparent displacement of external objects horizontally, momentary diplopia may result, but the eyes are capable of changing the convergence independently of the accommodation so as to regain single clear vision. If the prism is inserted so as to cause a vertical displacement, the diplopia which results cannot be overcome unless the displacement is very small ; the eyes have never been trained to make vertical readjustments relative to one another.

If the retinal images are made dissimilar, the stimulus to convergence no longer exists, and any defect in the association between convergence and accommodation can be investigated. These are known as defects in *muscle balance.* A subject looking through a Maddox rod (made up of a series of cylinders of red glass) in front of one eye, at a point source of light, will see the light as a streak with one eye, whilst with the other eye he will see it naturally. The white point and red streak are so dissimilar that there is no stimulus to visual fusion, and the amount of convergence of the subject's eyes is that which is inherent in the accommodation necessary to focus the light. If a concave lens is placed in front of the naked eye, a certain amount of accommodation will result, and with this will be associated additional convergence. Under these conditions the distance between the white spot and the red streak will change. Normal eyes should converge to exactly the right amount to correspond with accommodation for any distance, but it is usual to find some abnormality. Provided the eyes do not deviate vertically in their position of rest, the defect does not usually require correction.

Strabismus (squint) is a condition where it is obvious that the visual axes of the two eyes do not meet at the point of fixation. The condition may be due to paralysis of an eye muscle, often the external rectus, owing to the long exposed intracranial course of the nerve ; but it is also often found in hypermetropia. The chain of events leading to squint in hypermetropia is interesting. In order to focus objects, hypermetropes must accommodate more than

normal people, and with this is associated a stimulus to converge more than is necessary. Usually it is possible for the patient to counteract this urge to converge, but in cases where this is impossible the patient suppresses the image from one eye in order to avoid diplopia. This suppression of one retinal image becomes habitual, and being virtually blind it does not matter what position the eye takes up—it may become a squinting eye. Unless the hypermetropia is treated by glasses in early years so as to encourage binocular vision, the squinting eye becomes almost blind owing to disuse.

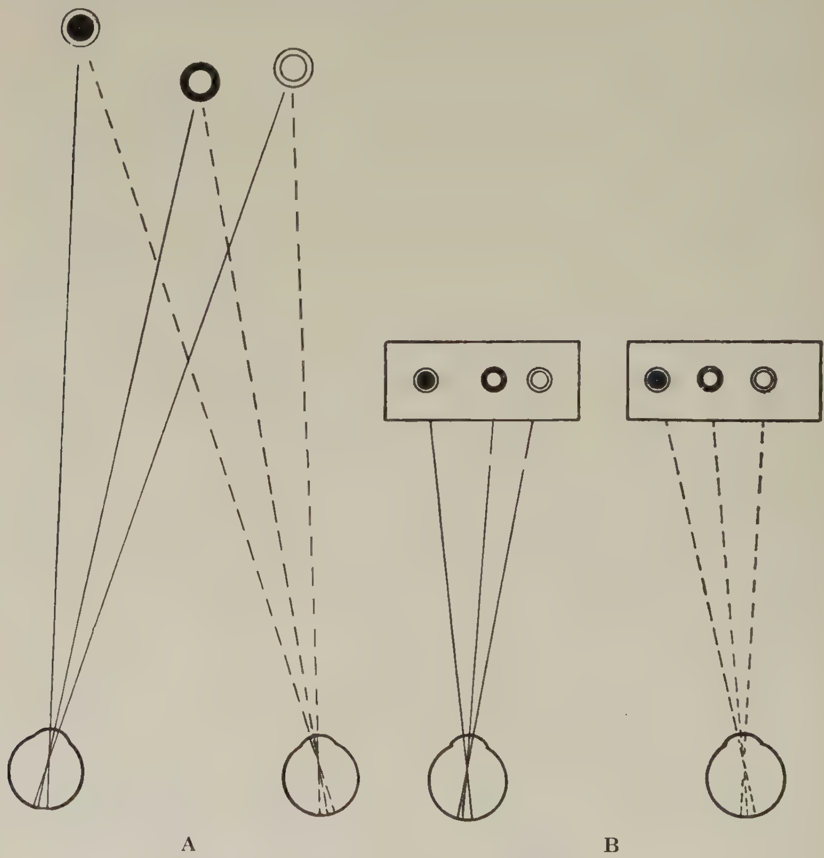


Fig. 18. 20. **Stereoscopic Vision.**

A shows the angles subtended to each eye when three circles, at different distances from the eyes, are viewed naturally. B shows a diagram placed in front of each eye reproducing these angles. This is the principle of the stereoscope, with which instrument the circles would be seen in relief.

Stereoscopic Vision. The images falling on the two retinæ are not quite similar, and advantage is taken of these differences in order to judge the relative distances of outside objects. By closing the eyes alternately it will be seen that where the differences between the two retinal images are great, one or more of the objects regarded will be seen double, but where the differences are only small the objects are

seen single, but acquire an additional appearance of solidity, by which their relative distances can be judged. Fig. 18. 20A shows the angles subtended at the two eyes by three circles at different distances from the eyes. If these objects are photographed, first from the position of the left eye, and then from that of the right eye, the two pictures which result can be arranged so that each is visible to the appropriate eye and not to the other (Fig. 18. 20B). The conditions of vision of the original objects have now been reproduced, and under these conditions the combined photographs have the same quality of depth as had the original objects. Owing to the urge to converge when accommodating for the photographs, most people require prisms to displace the images into a convenient position for fusion, but it is possible to " let the eyes go," so that the two parts of the figure coincide, in which case the circles will appear to be at different distances from the eyes.

Very small differences in angle are sufficient to make a judgment of relative distances ; stereoscopic vision is, in fact, one of the most exact visual judgments. It is, however, quite possible to get on perfectly well without it, and there are other ways of judging distance. The most important, perhaps, is the apparent relative movements of objects in the field of view when the head, or the whole person, is moved ; this is particularly valuable when one is in a moving vehicle. In addition, the relative sizes of recognisable objects, the overlapping of contours and shadows, and the tendency of distant objects to be hazy and to look blue, are all used separately or together. It is a common observation that in a hazy or misty atmosphere, objects are apt to seem far away and large, while in a very clear atmosphere they seem close to and small.

CHAPTER 19

HEARING AND SPEECH

HEARING is a sensation which is based on the activity of certain receptors which are sensitive to mechanical deformation. These are intimately associated with elaborate anatomical structures which ensure, first, that the receptors are activated by mechanical vibrations in the air around us—*i.e.* that they detect sound ; in consequence, the receptors detect, indirectly, mechanical vibrations in distant objects, with which we have no contact except through the intervening air. Secondly, the associated structures permit these vibrations to be analysed into patterns—in time more than in space—so that we have an auditory perception of the external world in addition to the visual perception based on information derived through our eyes. The ears, therefore, are distance receptors and hearing is a " special sense."

Of the various sounds that reach us from the external world, those made by our fellow human beings, in the form of speech, are of special significance. Speech is intimately connected with hearing. As a method of communication, its content resides in the nature of its constituent sounds ; and the ability of the listener to detect, analyse and differentiate these, one from another, is as important as the ability of the speaker to produce them.

The Physical Properties of Sound

If a tuning fork is struck the " prongs " of the fork are set into vibration, that is, they move backwards and forwards with great rapidity at one moment coming nearer to one another, at the next farther away. If, now, a small mirror is attached to one such prong, so that a beam of light reflected from it moves in sympathy with the movements of the prong, it will be possible, by arranging that this beam falls on to a moving strip of photographic film, to obtain a record of the movements of the prong over a given time. In other words, the record will constitute a graph showing the position of the prong (ordinates) against time (abscissæ). Such a graph is shown in Fig. 19. 1. The completed record is said to show the *wave-form* of the tuning fork vibrations, and the distance between the resting position of the fork and its position of maximum deflection gives a measure of the *amplitude* of swing of the fork ; where this distance is large, the vibrations are said to be of large amplitude, and where it is small, of small amplitude. It is easily demonstrable that a fork vibrating at greater amplitude will produce a louder sound ; the record from such a fork will show the same number of waves in a given time, but the amplitude of each wave will be greater. When the wave-form of a good tuning fork is examined in this way, it may be shown that the curve thus reproduced is almost exactly the same as

that obtained by plotting the sine of an angle against the angle itself. Such a wave is thus known as a *sine-wave*, and the vibrations of the fork which give rise to it as *sinusoidal.*

A similar record may be obtained without a mirror on the fork. When the prongs of the fork vibrate, air pressure changes are set up in their vicinity, and these in turn set up pressure changes near-by ; thus the sound travels through the air in the form of waves, the direction of wave motion being the same as that in which the sound is travelling, unlike waves on the surface of water, whose motion is at right angles

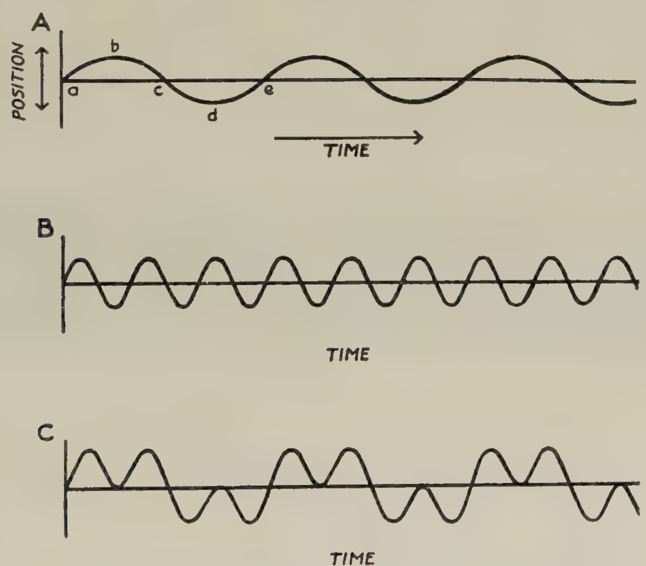


Fig. 19. 1. Wave-form of Vibrations of Tuning Forks.

A. a b c d e constitutes one complete cycle.
B. Frequency of fork three times that of A.
C Composite wave-form produced by sounding forks A and B simultaneously.

to their direction of travel ; these propagated pressure variations are perceived by man as sound. If a light diaphragm is arranged near a source of sound such as a fork, it will vibrate in sympathy with the changes in air pressure initiated by the source. By placing the mirror on this diaphragm, it will be possible to obtain a graphical record of its movements and thus of the air pressure changes. Such a record will be found to be exactly as Fig. 19. 1 (A), except that in this case the ordinates represent *air pressure changes.* The number of complete cycles, a b c d e, in one second of time is known as the *frequency* of the fork or note and is usually measured in *cycles per second.* The wire set into vibration by striking the note " Middle C " on a piano has a frequency of 256 cycles per second, that is, it performs 256 complete double

vibrations in one second.[1] The *pitch* of a note depends on its frequency; the higher the pitch the higher is the frequency. The octave above a given note has double the frequency, thus, the octave above "Middle C" has a frequency of 512, and the octave below it has a frequency of 128.

If, now, a fork of three times the frequency of the first fork be set into vibration, the result in diaphragm movements will be as shown in Fig. 19. 1 (B). Should forks of both of these frequencies be set going at the same time, the resultant wave will be that shown at (C). Thus, it is evident that the wave-form of the air vibrations set up by two forks sounding simultaneously is quite different from that obtained from either fork alone. Most of the sounds commonly met with in daily life have an infinitely more complicated wave-form than this, the *quality* of a sound depending on its wave-form. The wave-form of the vowel "ah" is shown in Fig. 19. 2. According to Fourier's theorem, however, just

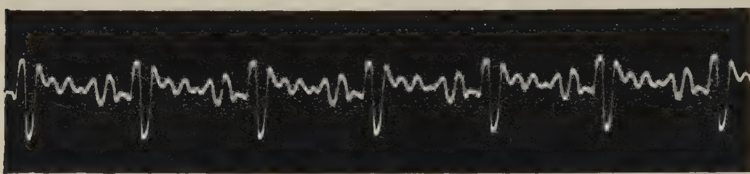


Fig. 19. 2. **Wave-form of an English Vowel of the type "ah."** **(D. B. Fry. from "Science and Speech.")**

as the relatively complicated wave-form of (C) is the result of adding together (A) and (B), so the wave-form of any periodic vibration, however complicated, may be resolved into a series of simple waves, consisting of a *fundamental* with the lowest frequency found, and a number of *overtones* or *harmonics*, whose frequencies are multiples of that of the fundamental. Each of these single component frequencies will be of the same wave-form as Fig. 19. 1, *i.e.* sinusoidal, and in the case of highly complicated sounds, the number of these components may be very great.

The analysis of a complex wave-form into its sinusoidal constituents in the ordinary way requires much tedious mathematical manipulation; but mechanical and electrical frequency analysers have been developed which have made possible the estimation of the frequency and power of the fundamental and harmonics of most sounds commonly met with in daily life. The frequency of these harmonics will bear a simple numerical relationship to that of the fundamental. They are called second, third, etc., harmonics, according as their frequencies are respectively twice, three times, etc., that of the fundamental or first harmonic. A sound whose wave-form is not strictly periodic will be found upon analysis to contain components whose frequencies are not simple multiples of that of the fundamental. These components are

[1] A frequency of 256 cycles for the note "Middle C" is traditional in physics. Pianos and other musical instruments are normally tuned to the international standard pitch (A = 440 cycles per second), corresponding, when adjusted for equal temperament, to a frequency of 261·6 cycles per second for middle C.

called *anharmonics.* " Noises," whose wave-form is highly irregular, contain many such constituents.

When a sound is emitted from a source, power is radiated in the form of pressure variations in the surrounding air : the *intensity* of the sound is the amount of power (in watts) which passes through unit area (1 square centimetre) at right angles to the direction of propagation. In practice, it is more convenient to measure the variations in air pressure (in dynes per square centimetre), by means of a suitable calibrated microphone, for example ; the sound power is proportional to the square of the pressure variation.

A just perceptible sensation may be produced by a sound whose intensity is somewhat less than 10^{-15} watt per square centimetre ; the variation in air pressure is then about 10^{-9} atmosphere, and the air molecules vibrate with an amplitude less than one-tenth of their own dimensions. In quite ordinary circumstances, however, we are exposed to sounds with intensities between a thousand times and a thousand-million times greater than this. It is more convenient, therefore, to avoid the use of very large figures by adopting a logarithmic scale (analogous to the *p*H scale for hydrogen ion concentration). The unit is called the *bel,* and the difference in bels between the intensity levels of two sounds is the common logarithm of the ratio of the two actual intensities : a 10-fold increase in the intensity of a sound is an increase of 1 bel in the intensity level, a 100-fold increase is an increase of 2 bels, and so on. For practical purposes, however, such a unit is inconveniently large, and the *decibel* is more often used. This is simply one-tenth of a bel, and represents an increase in intensity of roughly 25 per cent. (This is, as it happens, about the least change in intensity that can ordinarily be detected—see p. 578.)

Since the intensity (power) of a sound is proportional to the square of the variations in air pressure, the difference in intensity level between two sounds, in decibels, is given by twenty times the common logarithm of the ratio between the two sound pressures. (Since the pressure variations are alternately positive and negative, we define their magnitude in terms of the " root mean square (r.m.s.) pressure "; similarly, the voltage of an alternating electric supply is defined by the r.m.s., or " effective," voltage.)

It is to be noted that the decibel is a ratio and as such can have no meaning unless the standard from which it is measured is quoted. A number of different standards are commonly used for different purposes. One may use an arbitrarily chosen r.m.s. sound pressure, such as 0·0002 dyne per square centimetre, as conventionally used by acoustical engineers. But when measuring deficiencies in hearing, it is customary to use, as standard, the threshold intensity, at any frequency, of the average person with " normal " hearing, as given in Fig. 19. 8 below : the " hearing loss " of the patient is then the amount in decibels by which the intensity must be increased before the patient can just hear the sound.

Frequency, wave-form and intensity are physical properties of the

propagated waves of pressure change, whether they can be " heard " or not ; pitch, quality and loudness are of psychological significance, and describe characteristics of the sensation evoked by a sound wave. Pitch and quality are, for all practical purposes, determined uniquely by frequency and wave-form respectively. The pitch associated with a given frequency may, however, depend somewhat on the intensity, and the quality associated with a given wave-form may depend on both the intensity and the frequency of the fundamental component. The loudness of a sound of given intensity, on the other hand, may differ very considerably according to its component frequencies ; and the discrepancy is of sufficient importance to require special discussion in a later section.

The Structure of the Ear

The ear (Fig. 19. 3) is divided anatomically into three parts—the *outer ear*, the *middle ear*, and the *inner ear* or *labyrinth*. Each plays its part in the events leading up to the sensation of sound. The pinna and

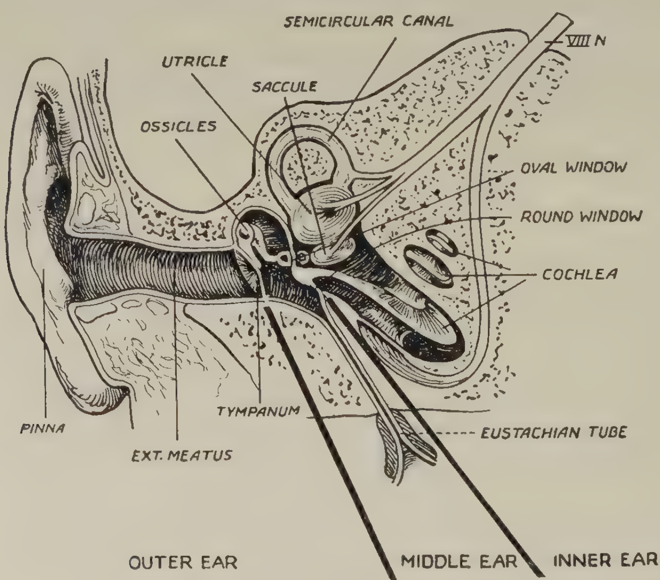


Fig. 19. 3. Diagram of the Ear.

external auditory meatus which comprise the outer ear are responsible for the collection of sound waves. The bony ossicles of the middle ear conduct the sound to the cochlea, a specialised part of the inner ear where the sound wave is analysed and the auditory nerve impulses are originated. Two other structures included anatomically in the inner ear, the vestibule and the semi-circular canals, are concerned with the sensations due to movement and orientation, as described in Chapter 17 (p. 513).

The *pinna* of most animals can be moved to the best position for the collection of the sound waves, but in man this organ is of little use and sounds can be heard almost as well without it. The *external auditory meatus* is protected against foreign bodies by hairs projecting outwards and from insects by a secretion of bitter wax. The meatus is 25 mm. long, and its end is blocked by the *tympanum* (ear drum), which is a diaphragm of connective tissue covered externally by modified skin continuous with that of the meatus, and internally by a mucous membrane continuous with that of the middle ear. The tympanum is conical in shape, with its concave surface facing outwards and downwards. It is supposed to be aperiodic, that is, it does not resonate in response to any particular frequency as do the diaphragms of most loud speakers. Attached to the apex of the cone on its inner side is the handle of the

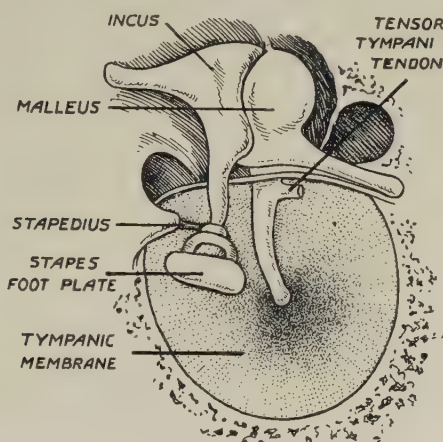


FIG. 19. 4. Semi-diagrammatic representation of the **Middle Ear.** (Modified from Beatty.)

malleus, the first of the three bony ossicles of the middle ear (Fig. 19. 4). Movements of the tympanum set up by the variations in air pressure due to the incoming sound waves are translated into movements of the malleus. Large variations of external air pressure are liable to rupture the tympanum, but this danger is minimised by the *Eustachian tube* (auditory tube) leading from the cavity of the middle ear to the naso-pharynx. It allows equalisation of air pressure on the two sides of the tympanum. There is a valve at the lower end of the auditory tube which is opened by the tensor palati muscle during the act of swallowing, in order to allow free communication between the ear and the pharynx. This valve is normally closed, so preventing one from hearing one's own speech unduly loudly—a painful phenomenon experienced by some patients in whom the muscle is inactive. Airmen and divers equalise the pressure on the two sides of the tympanum by swallowing occasionally.

Movements of the malleus are communicated to the second of the

middle ear ossicles, the *incus*, and from this to the *stapes*. There is a saddle-shaped articulation between the malleus and incus, and movements are communicated both by this and a spur of bone on the malleus. There is a ball-and-socket joint on the lower process of the incus with which the stapes articulates (*the incudo-stapedial junction*). The oval footplate of the stapes fits into the oval window of the inner ear, and is held in position by an annular ligament, which also prevents the escape of fluid.

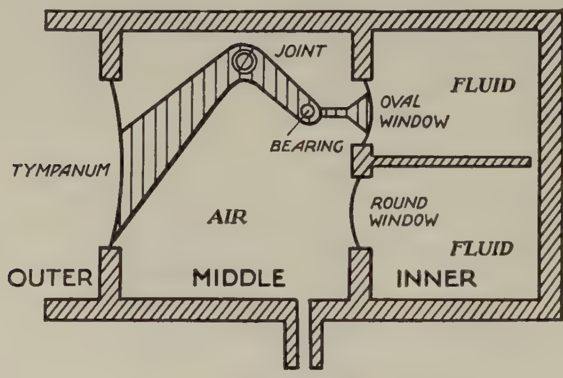


Fig. 19. 5. Model illustrating the method of transmission of vibrations from the outer to the inner ear. (Modified from Beatty.)

Attached to the ossicles are two small muscles, the *tensor tympani* which is attached to the malleus, and the *stapedius* which, as its name implies, is connected to the stapes. The stapedius is extremely minute, and is said to be the smallest muscle in the body. The function of these two muscles appears to be twofold ; first, the tensor tympani, owing to its tonus in the resting state, keeps the tympanic membrane taut, a necessary condition if its sensitivity to airborne vibration is to be maintained. Thus, if the tendon of this muscle is cut, the microphonic potentials of the ear fall by about one-fifth (see p. 585). The stapedius probably performs a similar function for the membrane of the oval window. Secondly, reflex contractions of both these muscles have been observed in dogs and rabbits, in response to loud sounds ; reflex contraction of the stapedius to such stimuli has been reported in man. Such contractions would be expected to have the effect of so damping the movements of the ossicles as to reduce the amplitude of vibration of the stapes footplate, and so of the oval window ; they would thus have a protective action against injury of the cochlea by very loud sounds.

The ossicles thus form a system of levers, as indicated very diagrammatically in Fig. 19. 5, and their function is to reduce the amplitude and increase the force of the vibrations imparted by the tympanum. The tympanic membrane is easily set into vibration by airborne sound waves ; the oval window, on the other hand, is heavily loaded by the cochlear fluids. It may be shown, both theoretically and experimentally, that in these circumstances power is transmitted more efficiently from the one to the other if there is a reduction in the amplitude of the vibrations, together with a corresponding increase in the force exerted. A

mechanism which has this effect may be seen on the old-fashioned acoustic gramophone. A short needle is attached to a lever, the other end of which is connected to the centre of a diaphragm : movements of the diaphragm set up movements of the air which pass to, and are amplified by, the horn. The pivot about which this lever moves is much nearer the needle end than the diaphragm end, so that a small movement of the needle produces a relatively large movement of the diaphragm. In the middle-ear conditions are reversed, and a large

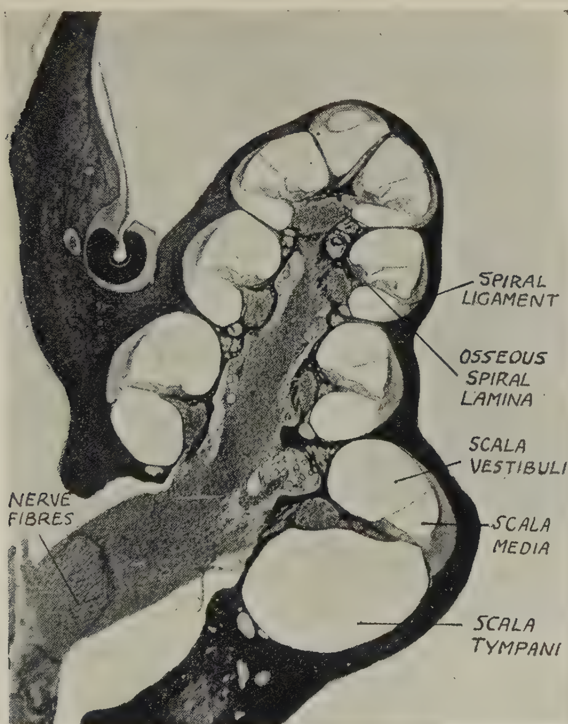


FIG. 19. 6. Section of the **Cochlea** of a guinea-pig, magnification × 7. (From a photograph kindly supplied by Dr. C. S. Hallpike.)

movement of the tympanum must provide a small movement of the oval window. In fact, the travel of the stapes footplate is only about one-third of that of the centre of the tympanic membrane ; the ossicles, therefore, appear to provide just such a reduction ratio as is necessary to ensure efficient transmission of airborne vibrations to the fluids within the cochlea.

In spite of the elaborate anatomical detail of the cochlea, its bare architectural outline is not difficult to understand. It is a cavity in the temporal bone shaped like a snail's shell ; from this it derives its name. Thus it resembles a tube coiled into the form of a conical helix,

round a central core of bone—the *modiolus.* The top of the tube is closed, but the bottom end contains in its bony wall two windows, each sealed by a membrane. The lower of these two is known as the *round window,* whilst the upper is called the *oval window* (Fig. 19. 5). A section made through the apex of the cochlea (Fig. 19. 6) reveals the cross-sections of the tube in outline, and the blood vessels and auditory nerve embedded in the modiolus. The lumen of the tube of the cochlea is divided into three compartments by two membranes and their attachments (Fig. 19. 7). The *basilar membrane* is stretched between two linear projections on the inside walls of the tube. The first of these projections is composed of bone, the *osseous spiral lamina,* and may be regarded as a projection of the wall of the modiolus into the lumen

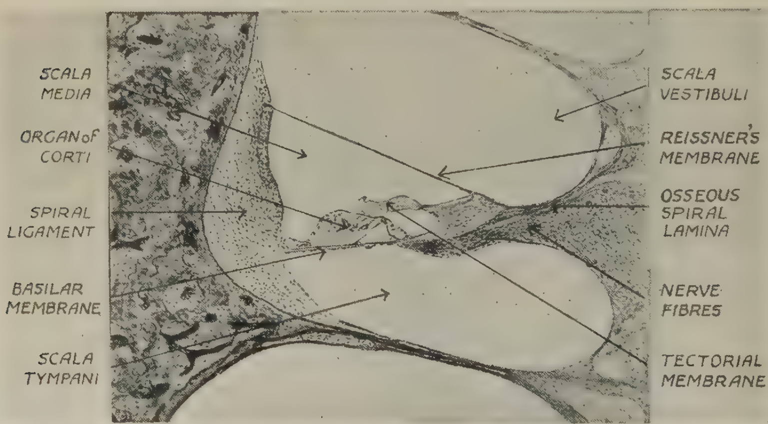


FIG. 19. 7. Section through the **Cochlea** of a dog showing the organ of Corti ; magnification × 75. (From a photograph kindly supplied by Dr. C. S. Hallpike.)

of the tube. It therefore forms a ridge running from the apex to the base of the cochlea and resembles the marking of a screw. On the opposite side of the tube, also forming a projecting ridge along its whole length, is the *spiral ligament* composed of fibrous tissue. Reissner's membrane arises from the inner side of the *Organ of Corti,* and runs upwards and outwards at an angle to become attached to the wall of the tube above the spiral ligament. The space enclosed between these two membranes is known as the *scala media,* and it contains endolymph. The space above Reissner's membrane contains perilymph, and is known as the *scala vestibuli* ; the space below the basilar membrane also contains perilymph, and is known as the *scala tympani.* These two spaces are continuous at the apex of the cochlea, through an opening known as the *helicotrema.* The oval window is between the scala vestibuli and the middle ear cavity, the round window between the scala tympani and the middle ear. The escape of fluid from the inner ear is prevented at both windows by a membranous covering.

It will now be necessary to consider how the movements of the middle ear ossicles are transmitted to the basilar membrane, movements of which stimulate the auditory nerve. When the foot of the stapes is displaced inwards into the oval window, pressure is applied to the perilymph, tending to compress it. Since this fluid in its bony casing is incompressible, movements in it must find some outlet. This is provided by the round window membrane, which bulges into the air cavity of the middle ear, the movements being, therefore, a counterpart of those at the oval window. Slow displacements of perilymph in the scala vestibuli could be transferred to the perilymph in the scala tympani by the helicotrema ; the movements are so rapid, however, that this cannot take place. The most important channel for accommodating the displacements of fluid is, therefore, by means of the scala media and its two limiting membranes. A displacement of the stapes inwards causes a downward movement of the basilar membrane, and a displacement outwards causes an upward movement. These movements of the basilar membrane are intimately connected with the perception of sound.

The basilar membrane and its associated structures, in particular the organ of Corti, are shown in Fig. 19. 7. The *basilar membrane* itself is composed of fibrous tissue lined on its under side with connective tissue ; its length from base to apex of the cochlea is about 32 mm. ; its breadth at the base is about 0·17 mm., and at the apex about 0·47 mm. The organ of Corti is situated on the inner and upper aspect of the basilar membrane and forms a continuous chain from apex to base. Each segment of the organ of Corti consists of two pillars forming an arch of Corti which supports the inner and outer hair cells. Hairs protrude from the upper extremity of each cell, and are embedded in the tectorial membrane, a flap of connective tissue with its internal edge attached, forming a roof to the organ of Corti. Movements of the basilar membrane, and thus of the organ of Corti, result in deformation of the hairs, and impulses are initiated in the nerve fibres which are connected with the hair cells ; these cells are thus the primary receptors for sound. The nerve fibres are derived from ganglion cells in the *spiral ganglion*, just within the osseous spiral lamina, and leave the cochlea in the cochlear nerve which lies in the hollow centre of the bony modiolus.

The movements of the oval window follow almost exactly the wave-form of the incoming sound. But if the amplitude of the sound is large, imperfections of the tympanum and middle-ear ossicles may introduce some *wave-form distortion*. The effect is to add harmonics of the fundamental tones present in the air, or to increase the amplitude of those already present. If a pure tone of frequency n c/s is impressed on the ear drum, the vibrations of the cochlear fluids will contain notes of frequency n (the fundamental), $2n$, $3n$, etc., c/sec. The *subjective harmonics*, introduced by the ear itself, affect the quality of the sound, but as a rule it is only when the sound is both loud and low in pitch that the amplitude is great enough for them to be apparent. Even so, the subjective harmonics are likely to be small compared with those introduced by the instrument or equipment which makes the sound, unless this is very specially designed.

Auditory Acuity

When the intensity of a sound is continuously decreased, it reaches a value where no sensation of sound is produced. The smallest intensity of sound required to produce a sensation is said to be on the *threshold of audibility*. As is shown in Fig. 19. 8, the physical intensity (sound pressure level) of a threshold sound varies very considerably with its frequency. The ear is not uniformly sensitive throughout the range of frequencies : it requires a smaller intensity of sound at 1,000 to 4,000 cycles per second to produce a just perceptible sensation than it does at frequencies higher or lower than these ; at 100 c/sec and at 10,000 c/sec the threshold intensity is nearly 1,000 times greater than it is at 3,000 c/sec.

The curve of threshold intensities given in Fig. 19. 8 represents the average values obtained from a group of young people, aged 18–25, using one ear only and applying the sound through an earphone. This is the method ordinarily used for testing deficiencies in hearing. But we usually hear sounds in the open air, using both ears : in such " free field " conditions, the threshold of audibility is some 5 to 10 decibels (depending on the frequency) below the values plotted in Fig. 19. 8.

If the intensity is adequate, a pure tone may be heard, and recognised as such, if its frequency lies between some 20 and some 20,000 c/sec. Many kinds of animal, however, such as dogs, can hear notes with frequencies well above 20,000 c/sec, which are inaudible to man. Beyond these limits of frequency, a " sound," if sufficiently intense, may be felt by parts of the body other than the ear, but it will not, strictly speaking, be " heard." Within these limits of frequency, a very intense sound is not only heard, but also felt by the ear, often a somewhat painful sensation ; it is then said to be on the *threshold of feeling*.

Loudness. The loudness of a pure tone may be measured in terms of the physical intensity of the tone, expressed in decibels above the physical intensity of the same tone when it can just be heard—*i.e.* at the threshold of audibility—as given, for example, by the lowermost curve in Fig. 19. 8. Tones of different frequency, but of the same physical intensity, therefore, will not necessarily have the same intensity level above threshold, and are not necessarily of the same loudness. We cannot, therefore, measure the loudness of a complex sound in terms merely of its physical intensity, since in general we do not know what its physical intensity would be if it were reduced in loudness to the threshold value ; this will depend on the relative magnitudes of its component frequencies. This difficulty is avoided by using a pure tone of 1,000 cycles per second as a reference standard, and adjusting its intensity until it is judged, by a person with " normal " hearing (or a group of such persons), to have the same loudness as the sound or noise under investigation. The intensity of the 1,000 c/sec., tone is measured in decibels above a reference sound pressure of 0·0002 dyne per square centimetre. This then becomes a measure of the subjective loudness

of a complex noise, for example, and is expressed in terms of a unit called the *phon*. For all frequencies other than 1,000 c/sec, and for complex noises, the subjective loudness (measured in phons) differs, often very considerably, from the physical intensity (measured in decibels above the reference pressure). This is indicated by the three upper curves in Fig. 19. 8, each of which represents the relation between physical intensity and frequency for pure tones, which have the same subjective loudness of 40 phon, 80 phon and 120 phon, respectively.

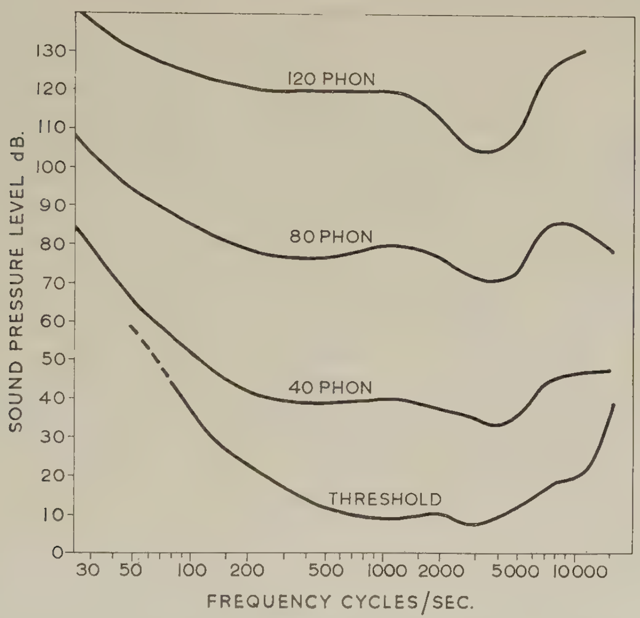


FIG. 19. 8. The Variation of **Intensity** with **Frequency** (a) for just detectable (**threshold**) tones and (b) for tones of equal **loudness** (at three different levels).

The intensity is measured in terms of the sound pressure level, in decibels above a pressure of 0·0002 dyne per square centimetre. (Robinson and Dadson, from Kaye and Laby, "Tables of Physical and Chemical Constants" 11th edition, 1957.)

On the whole, the greater is the intensity of a sound, the less does its loudness vary with its frequency. It is to be noted that since the phon is measured in decibels above an arbitrary reference level, a sound at the threshold of audibility will not necessarily have a loudness of 0 phon.

In practice, loudness measurements are often made, not by subjective judgment, but by using a sound intensity meter (*i.e.* an instrument measuring sound power directly) whose frequency characteristic can be adjusted to match that of the ear. Such instruments are usually calibrated in decibels above a reference sound pressure level ; loudness is thus very commonly expressed in " decibels," although the phon is

the correct unit. Approximate values of the loudness of certain noises are given in Table 19. 1.

TABLE 19. 1

		Loudness in phon
Faintest audible sound (average normal person)		4
Whisper	about	20
Quiet street	,,	40
Ordinary conversation	,,	60
Busy street	,,	80
Noisy engine room	,,	110
Painful sound	,,	130

Difference, or Discrimination, Thresholds. A normal man is capable of detecting quite small differences between the *frequencies* of two tones. The minimum perceptible difference (*absolute* difference threshold) is about 3 c/sec, independent of the frequency up to about 500 c/sec, and rising more or less in proportion to the frequency above about 1,000 c/sec. Expressed as a percentage of the frequency, the *relative* pitch discrimination threshold falls with increase in frequency up to about 1,000 c/sec, and then remains more or less constant, as is shown in Fig. 19. 9. Up to a limit, the ability to discriminate between sounds of different frequency improves with increase in the intensity of the sound, as is indicated by the three curves in Fig. 19. 9. If the pitch is fairly high, therefore (frequency greater than 1,000 c/sec), a difference of about 1/20 semitone can be detected : if it is low (say, 100 c/sec) about 1/2 semitone. Over the whole range of audible frequencies there are about 1,600 perceptible changes in pitch.

The acuity of the ear for differences in sound *intensity* is much smaller than its acuity for differences in frequency, the fractional increase which can just be detected under the best conditions being between 5 and 10 per cent. Intensity discrimination is not so good for low notes or low intensities ; at a frequency of 60 c/sec and at a moderate intensity, a change of 20 per cent. is just perceptible, whilst at low intensities an increase of as much as 200 per cent. is necessary.

Masking. It is a common experience that the presence of one sound reduces our ability to hear other weaker sounds, which are then said to be masked. Suppose that we are listening to some particular sound whose intensity is at about the threshold value, and other sounds or noises then begin. In order just to hear the sound that we are interested in, we should have to increase its intensity to a value which is about 10 to 30 decibels below that of the interfering sounds : the threshold of audibility is thus raised by an amount which increases with increase in the intensity of the masking sounds. But much depends on the frequencies of the sounds considered ; a loud sound of any particular frequency masks others with a higher frequency much more than it does those with a lower frequency. If we study the masking effect of a pure tone, we find that if its intensity is not very large (say

up to about 40 dB above threshold), only notes immediately above and below it in frequency are appreciably affected ; the threshold of audibility of these is raised to a value which is about 10 dB less than the intensity of the masking tone. Notes which are an octave or more away on each side are not affected. If the intensity of the masking tone is raised, however (say to 80 or 100 dB above threshold), the effect spreads increasingly to notes of all frequencies above it, whilst notes of much lower frequency, an octave or more below it, remain unaffected ; the

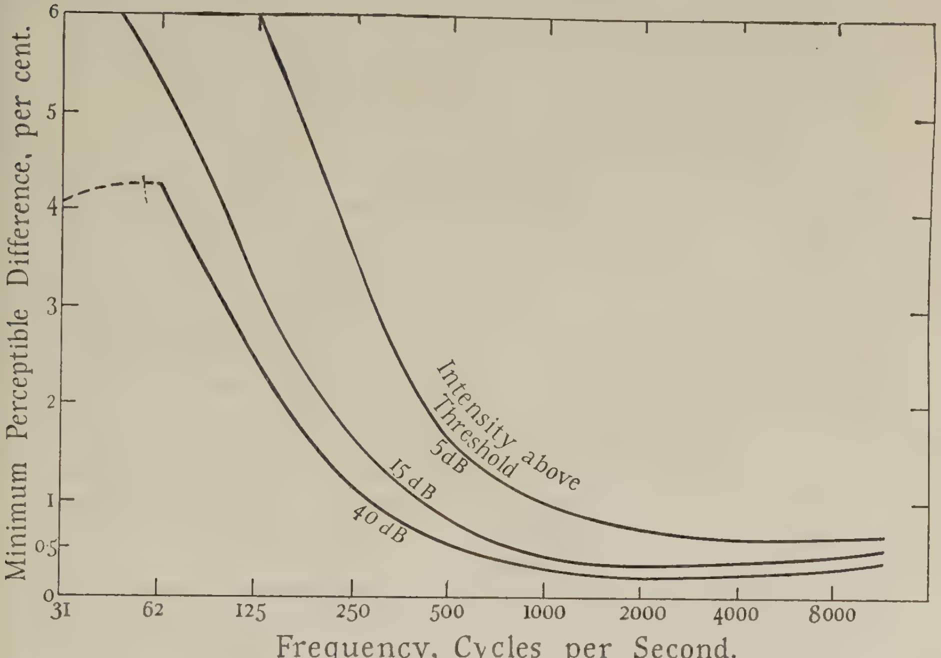


FIG. 19. 9. The **Minimum Differences in Frequency** which can be detected at various frequencies and at various intensities above the threshold (expressed as a percentage of the frequency of the stimulus). At intensities greater than 40 dB, the changes in sensitivity are small. The values given are for listening with one ear ; if both ears are used, the minimum perceptible differences are smaller. (From data by Shower and Biddulph.)

threshold of audibility of all the masked tones is now raised to a value which is about 30 dB below the intensity of the masking tone.

These masking effects are to be expected from the properties of the cochlea as discovered by direct observation and by electrophysiological studies (see p. 583, below) ; they affect, also, the intelligibility of speech in the presence of much background noise (p. 588) ; and they affect the relation between the quality of a complex sound and its wave-form. Suppose, for example, that we produce a complex sound consisting of a fundamental frequency A together with notes of progressively

increasing frequency a, b, c, etc. If the intensity is small, all the constituents may be heard : if the intensity is doubled, A may mask a ; if if it is redoubled, A may mask both a and b ; and so on until nothing will be heard but the single note of frequency A, which therefore gains in prominence as the sound is amplified.

The Localisation of Sound. The ability to recognise the direction from which a sound proceeds is poorly developed in man. Under the best conditions it is rarely possible to localise the direction of a sound more accurately than within 10 degrees in the horizontal plane and considerably more in the vertical plane. Man's ability to localise sound is due to two factors. First, sound has a finite velocity, so that unless a source of sound is equidistant from the two ears, a given vibration will not arrive at both ears simultaneously : a sound from a source at an angle of 45 degrees from the sagittal plane, for example, will reach one ear approximately 0·4 millisecond before it reaches the other. If the two ear pieces of a set of headphones are wired separately, so that clicks may be presented to the two ears with a small time difference between them, it is found that the direction from which the sound appears to come varies with this interval, provided that it is not greater than about 1 millisecond. Secondly, for high frequency tones, above 3,000 c/sec, there will be a difference of intensity at the two ears due to the " shadow " effect of the head ; the sound is localised to the side where it is loudest. Both methods are probably used together for localising the most common sounds, which contain components of many different frequencies.

Many animals have another way of localising sounds : they rotate their pinnæ until the apparent intensity is greatest, when the pinnæ will be facing the source of sound.

Auditory Fatigue. If the ear be subjected to prolonged loud sounds, is suffers a transient loss of sensitivity, *i.e.* it becomes fatigued. This means that the intensity of a note, if it is just to be perceived, must be made greater after such stimulation than before. A sufficiently intense stimulus, applied to a single ear only, causes a loss of sensitivity in both ears ; that is to say, the fatigue is binaural. The loss in the nominally unstimulated ear is not produced by the residue of sound reaching it (due to bone conduction, slight leaks, etc.), as is shown by the following experiment. Suppose a sound of intensity only 70 decibels above the threshold be applied to the right ear, then no loss of acuity will be found in either right or left ears. If, however, the intensity of the fatiguing sound is raised, although still applied only to the right ear, the left ear, as well as the right, will suffer a loss of sensitivity. Yet the residue of sound reaching the left ear is at least 60 decibels below that in the right ear, and thus can exert no fatiguing effect on it. It follows, therefore, that stimulation of one ear, if sufficiently intense, lowers the sensitivity of both ears ; this effect is probably central. The loss in the stimulated ear, however, is somewhat greater than that in the unstimulated one. The greater fatigue of the stimulated ear is presumably due to a loss of sensitivity in the peripheral mechanism.

In general, tones of frequency above 1,000 c/sec produce much more fatigue than do tones of frequency below this.

Deafness. The tests for deafness normally employed are the spoken voice, whisper, the tick of a watch, and tuning forks of various frequencies. Such tests are inaccurate, but in experienced hands can give a fair indication of the degree of loss of hearing. A more accurate instrument, with which the sensitivity over almost the whole auditory spectrum may be tested, frequency by frequency, is the *audiometer*. In this, pure tones produced electrically are led to the ear by a special telephone ; the intensity of the tone may be altered at will, and that point at which the patient no longer hears the sound is readily estimated.

The threshold of audibility, for tones of 10,000 cycles per second and above increases rapidly with increasing age, even in " normal " people, who have no particular disorder of hearing ; there is usually, also, a smaller loss of sensitivity to frequencies down to 1,000 c/sec. More serious deafness may result from several types of disorder.

(*a*) *External Ear Obstruction.* The external auditory meatus is obstructed by a wax plug, by dirt, or by inflammation and swelling of the meatal wall (otitis externa).

(*b*) *Middle Ear Disease.* The ossicles are prevented from functioning properly, a condition which frequently follows a nasal catarrh, and starts as an inflammation of the middle ear (otitis media), the infection entering via the Eustachian tube. Later, a pathological condition of the bone round the inner ear may develop (*otosclerosis*). This mainly results in the formation of new bone in the neighbourhood of the oval window which causes fixation of the stapes footplate to the bony capsule, with consequent loss of hearing. In the later stages, the organ of Corti often shows degeneration.

(*c*) *Inner Ear Disease.* This usually involves loss of function of the organ of Corti or of the cochlear nerve, and is thus sometimes referred to as " nerve deafness." It may be due to :—

(1) *Injuries to the Inner Ear*, such as boilermakers' deafness (Chronic Labyrinthine Concussion) ;

(2) *Diseases of the Inner Ear and Auditory Nerve*, due to local hæmorrhages or inflammation or to general disease (*e.g.* syphilis, malaria, rheumatism). The commonest cause of inner ear disease, however, is middle ear disease.

External and middle ear deafness may often be distinguished from inner ear (nerve) deafness by making use of the phenomenon of " bone-conduction." If the auditory meatus be carefully plugged, a tuning fork will be heard without difficulty if its stem is placed on the bones of the head (*e.g.* the mastoid bone). The vibrations from the fork set up similar vibrations in the cochlea, transmission taking place through the bones of the skull. It is clear, therefore, that if air-conducted sound is not heard, whilst bone conducted vibrations are readily audible, loss of middle ear function must be suspected. This diagnosis may often be strengthened by inspection of the tympanic membrane, which, in most cases of middle ear disease, presents an abnormal appearance.

Deaf Aids take many forms, but all depend for their action upon raising the intensity of the received sound above normal. This is done electrically by means of a microphone, amplifier, and headphone. With such an instrument the relative intensity of high and low tones may be altered at will, thus compensating for losses over a specific range of frequency. It is difficult, however, to avoid some unwanted distortion, and for this, and other reasons, they are not always entirely satisfactory.

The Auditory Receptor System

The auditory system can detect differences in frequency of an incoming sound of less than 5 parts in 1,000 over much of the audible range, and simultaneously it can detect differences in intensity of about 10 per cent. These facts must be explained.

The air-borne sound waves which enter the ear are transmitted through the oval window to the fluids of the scala vestibuli, unchanged in frequency but reduced in amplitude, as has already been described. The displacement of this fluid causes movement of the basilar membrane and thus displacement of the fluid in the scala tympani and movement of the round window. The organ of Corti, with its associated hair cells, is constrained to move with the basilar membrane. The first step in the analysis of the mode of action of the auditory receptor system, therefore, is the study of the way in which the motion of the whole structure of membrane and hair cells depends on the frequency of the vibration that is impressed on it. This has been done by von Békésy by direct observation of the basilar membrane of the fresh human cadaver, after grinding away the bone over the apex of the cochlea. He found that vibrations of low frequency (less than a few hundred cycles per second) cause the whole membrane to move, the amplitude being greatest near the apical end of the membrane, and becoming gradually smaller towards the basal end. At a frequency of 50 c/sec, there was a position of maximum amplitude a few millimetres from the helicotrema, and as the frequency was increased, the position of this maximum moved towards the basal end of the membrane (Fig. 19. 10) : at each frequency, the amplitude fell off rapidly towards the apical end and slowly towards the basal end. If the frequency was 200 c/sec or more, the apical end did not move at all ; and at a frequency of 3,000 c/sec, movement was seen only in a short length of membrane close to the oval window.

If the displacement of the basilar membrane is a direct precursor of the excitation of the receptors, one would expect the pattern of varying amplitude with distance from the oval window to be reflected in the behaviour of the primary receptor units. It is very difficult to record from such units, but it has been done by Tasaki. When he recorded from a primary unit, selected at random, he obtained a result such as that shown in Fig. 19. 11. With a stimulus of very low intensity, he obtained a response only at one frequency, presumably that which caused the largest displacement at that particular point at which the receptors were located. With larger stimuli, he obtained responses from greater and greater ranges of frequency, the extension of the range

being mainly towards the lower frequencies. This is what would be expected from the mechanical evidence, since every lower frequency would displace the basilar membrane from the oval window up to some point beyond the receptor from which the recording was made ; higher frequencies, which displace only a short length of the membrane would, on the other hand, leave the receptor in a quiescent part of the membrane. (Compare in Fig. 19. 10, for example, the displacement at different frequencies of a point on the membrane 25 mm. from the stapes.) The shape of the curve in Fig. 19. 11 is, in fact, entirely in accordance with the direct observations on the displacement of different parts of the membrane at different frequencies. Thus we see that the

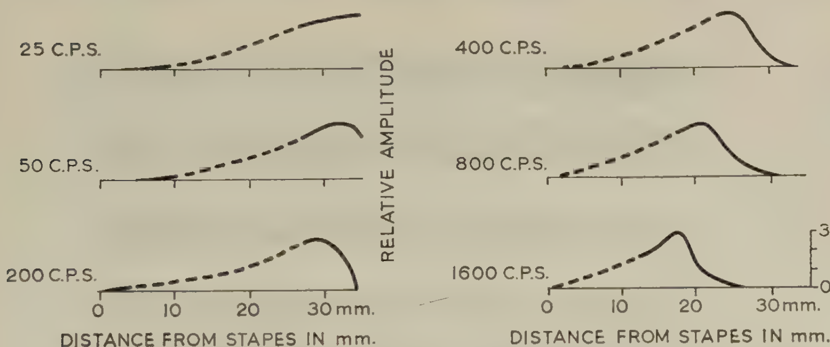


FIG. 19. 10. The **Displacement** of the **Basilar Membrane** in response to Vibrations of different Frequency.

Vibrations of constant amplitude but with different frequencies were impressed on the fluid within the cochlea, and the amplitude of the oscillatory displacement of the basilar membrane was measured at different distances from the stapes (basal end of the cochlea). At low frequencies, the amplitude is greatest at the apical end of the membrane : as the frequency is increased, the position of maximum amplitude shifts towards the basal end.

In order to observe the more basal parts of the membrane, the uppermost turns of the cochlear spiral had to be removed. This may be expected to affect the displacement of the parts observed at those frequencies at which the parts removed would themselves have undergone displacement : these observations are represented by the broken lines. Control observations showed, however, that in the conditions of measurement the error was quite small. (v. Békésy, J. Acoust. Soc. Amer. 1949.)

greater the intensity of a sound, the greater the length of the membrane in which receptors will be excited ; but that the distribution of these excited receptors along the length of the membrane will be different according to the frequency of the sound.

The observed displacement of the basilar membrane by sounds of varying frequency and intensity are in accordance with what might be expected from the phenomenon of masking. A masking tone, particularly if of high intensity, will cause the whole basilar membrane between the oval window and its point of maximum displacement to vibrate : thus those parts of the membrane associated with frequencies higher than that of the masking tone will be activated and masked.

Frequencies lower than the masking tone are able to displace the basilar membrane in the inactive regions nearer the apex.

It is clear, therefore, that no single receptor unit in the cochlea can transmit enough information to distinguish both frequency and intensity, since it will respond in the same way to a low intensity at its optimum frequency and to a higher intensity at some other frequency. The same problem is found in the skin, where a single receptor unit will respond in the same way to a small force in the centre of its receptive field as to a large force further away. The required amount of information to make all the necessary distinctions is provided in both situations by the use of large numbers of units each placed in a slightly different

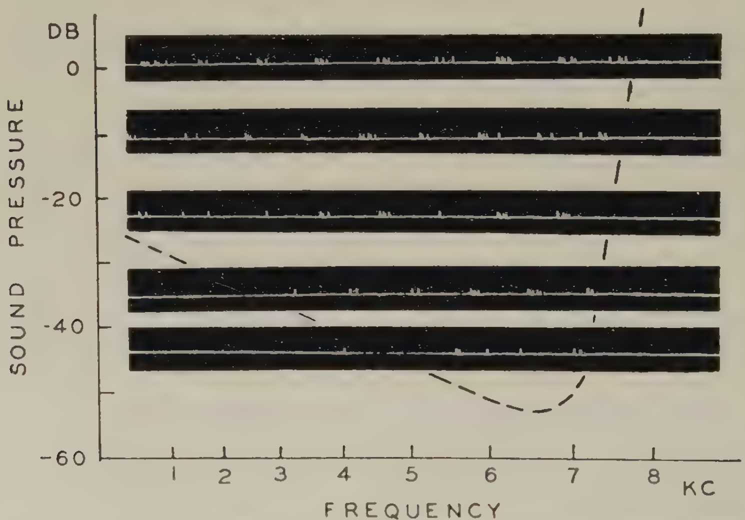


Fig. 19. 11. Action Potentials from a single fibre in the Auditory Nerve of a guinea-pig.

Each line, from left to right, gives the responses to tones of short duration (tone pips) in an ascending series of frequencies, as given in the scale at the bottom, at a constant value of the intensity. The intensity is greatest in the top line, and becomes progressively smaller in the lower lines. The broken line shows the relation between the threshold intensity and the frequency, for this particular receptor unit. The sensitivity is greatest between 6,000 and 7,000 c/sec., falls off gradually as the frequency is reduced and rapidly as the frequency is increased. (Tasaki, 1954, *J. Neurophysiol.*)

position. It is the task of the analysing apparatus of the central nervous system to work out the answer. How this is done is not yet understood, though it is already probable that inhibitory interactions between the receptor units play an important part in the auditory system, as they do in the reception of visual signals by the eye.

The idea that the cochlea played an essential part in frequency discrimination by causing different receptors to fire at different frequencies of vibration is much older than the evidence given in the last paragraph. The first of

these *place theories*, as they are called, was put forward by Helmholtz: this was his *resonance theory*. He postulated that the basilar membrane and the organ of Corti behaved as a set of strings, each of which was set to resonate at a different frequency. The theory as he proposed it is not consistent with modern findings, but the important point in his idea is that the displacements of different parts of the basilar membrane are different at different frequencies.

The displacement of the basilar membrane is a precursor of the initiation of nerve impulses, but is not the immediate cause of the excitation. In Chapter 17 it was pointed out that in most receptors a change of electrical potential, which is graded in amplitude according to the strength of the physical or chemical stimulus, is the immediate cause of the initiation of the nerve impulses. In single receptors the relationship between the stimulus and this receptor potential, and between the receptor potential and the nerve impulse, is fairly clear; but in the cochlea the situation is in some respects both more complex and more difficult to analyse. This is because it is not easy to investigate the behaviour of single receptors, although potentials developed across the whole organ have been thoroughly investigated. Potential changes are developed between the scala tympani and the scala media when sound enters the ear, and these changes are of the same frequency and phase as the sound stimulus, and are proportional to it in amplitude. These potentials are distributed along the cochlea in a way that is consistent with the pattern of displacement that has already been mentioned; that is to say, they can be found in response to high frequencies mainly near the oval window, and in response to low frequencies mainly at the apical end. But since the potential changes themselves spread along the cochlea, owing to its electrical properties, it is not easy to locate the sources accurately. These graded changes of electrical potential were first described by Wever and Bray and are termed *microphonic potentials*.

The reason for the name and also an indication of the accuracy with which the microphonic potentials follow the sound waves, can easily be demonstrated in the following way. If an electrode is placed on the round window of an experimental animal and another on, say, the neck, microphonic potentials of all frequencies can be picked up. If this output is amplified and put into a loud speaker, it is found that the system will reproduce voice and other sounds quite accurately. In other words, the cochlea used in this way acts as a microphone.

The basilar membrane lies between the scala tympani and the scala media, and thus has perilymph on one side and endolymph on the other. Now the composition of perilymph, as remarked in Chapter 8, is that of the extracellular fluid in general, the sodium ion concentration being large compared with the potassium ion concentration. In the endolymph, however, the potassium ion concentration is greater than the sodium ion concentration, so that its composition is more like that of the intracellular fluid. There is an electrical potential difference between the endolymph and the perilymph and, on the arrival of sound stimuli, this *endolymphatic potential* varies about its "resting" value in accordance with the variations in sound pressure. The microphonic

potentials, therefore, are modulations of the endolymphatic potential; their power is derived only to a small extent from the incoming sound waves, and chiefly from the metabolic reactions which create and maintain the endolymphatic potential. This is a good example of the process of "power amplification" which is common, if not universal, in receptor systems, as pointed out in Chapter 17.

The source of the endolymphatic potential is not definitely known. It is opposite in sign to the resting potential between the interior of a cell and the extracellular fluid, the endolymph being electrically *positive* to the perilymph by about 80 mV.

At any particular level in the cochlea, the greatest potential field (or gradient) due to the microphonic potential has been found to be in the region of the hair cells and of the associated nerve fibres. Furthermore, it has been found that nerve impulses are set up on each cycle of the microphonic potential, provided that the frequency is not too high. At low frequencies, with sufficiently large stimuli, there may be a burst of impulses during each cycle; but as the frequency gets higher, there is time for one impulse only on each wave. When the interval between successive cycles becomes comparable with the refractory period of the nerve fibres (both relative and absolute), which may range from 3 msec down to 0·5 msec—*i.e.* when the incoming frequency reaches 300 to 2,000 c/sec—impulses will not appear on every cycle of the microphonic potential, the number missed depending on the stimulus strength. These observations are important for two reasons: first, they suggest strongly that the microphonic potential is associated with the initiation of nerve impulses, even if it is not the immediate cause; and secondly, that the timing of the impulses is closely related, at certain frequencies at least, to the intervals between the peaks of the sound wave. If recordings are made of action potentials in the auditory nerve, it is found that the discharge is intermittent, and related to the phase of the sound wave, up to frequencies of about 4,000 c/sec. No single fibre can be activated at these frequencies. But since the triggering of an impulse in any given fibre is related to the phase of the microphonic potential, any fibre which is ready to fire (after recovering from the refractory state) will not do so until the appropriate moment in the next cycle: in a whole population of fibres there will always be some which fire at each cycle, although any individual fibre will fire on every second, third or fourth cycle, as the case may be. There is thus a temporal pattern of nerve activity, as well as a spatial pattern, which is related to the frequency of the sound wave; how, or even whether, the nervous system makes use of these clues is not established.

There is one type of evidence which shows that the position of the displacement of the basilar membrane is a necessary clue in the determination by the brain of the frequency of a sound wave. High intensity sounds can, in experimental animals and through industrial injury in man (as in "boilermakers' disease," for example) cause damage to a localised part of the basilar membrane. That such damage can be seen *post mortem* is merely confirmation of what is known of the mechanical

behaviour of the cochlea. It does not tell us whether or not the brain finds the analysis so made essential. There is, however, some evidence that such injuries are accompanied by an inability to appreciate sounds that have a frequency which would cause a maximum displacement in the injured region of the basilar membrane.

SPEECH

The sounds of speech are produced, first, by the movement of air exhaled from the lungs past the vocal cords in the larynx, which behave like the reeds of wind instruments and are set into vibration ; the frequency depending on the amount to which they are stretched, and put under tension, by the action of the laryngeal muscles. The resulting sound contains many different frequencies besides the fundamental, and certain of these are reinforced by resonance in the cavities of the throat, mouth and nose ; adjustment of the size and shape of these cavities, individually, controls the nature of the speech sound, particularly of the vowels, which is emitted. Secondly, the movements of air over sharp edges such as the teeth, and through narrow gaps formed between the tongue and the soft palate and between the lips, produce " hissing " sounds which give rise to some of the consonants, those known as " fricative," such as s, f, v and th. In addition, the way in which the sounds are started and stopped, by the action of the tongue and the lips, gives rise to the " stop " consonants, such as b, p, g and k. Adjustment of the vocal cords alters the general pitch of the speech ; but the distinction between one kind of speech sound and another depends on the action of the muscles of the throat, cheeks, tongue and lips. The accurate control of these various muscles is, to a large extent, reflex. We listen to our own speech as we make it, and adjust the sound so as to imitate that desired. If a person's speech is presented to his ears, not instantaneously as ordinarily occurs, but after a delay of a few tenths of a second, he becomes so confused that further speech becomes almost impossible.

This may be done by recording his speech on a tape from which it is picked up again after a short delay and played loudly into his ears through telephone receivers ; the direct and immediate transmission from his mouth to his ears is thus overpowered by the delayed transmission.

Deaf children do not hear other persons' speech, or their own, and do not ordinarily learn to speak spontaneously. They will do so if they can be given adequate hearing aids, and can be taught by special methods. Their speech, however, though comprehensible, usually has a peculiar quality.

Speech sounds have extremely complex wave-forms—that illustrated in Fig. 19. 2 being a relatively simple example—and may contain components with all frequencies between about 800 and 8,000 c/sec. Most of the power (about 1/3) is in frequencies between 250 and 500 c/sec, and these, therefore, largely determine the loudness of the speech. But the distinction between one kind of speech sound and another, and thus the intelligibility of the speech, is determined by the higher frequency components, chiefly those between 500 and 5,000 c/sec. Provided

that these are all present in their proper amplitudes in the speech sounds, and are adequately heard by the listener, there is no serious difficulty in recognising all the various speech sounds that may occur. Ordinary speech, in a familiar language, can be understood even when the only frequencies heard are those between 1,000 and 3,000 c/sec: but this is largely because in these circumstances we do not need to hear and identify every speech sound, or even every word ; those missed can be supplied from the context. Since the higher frequencies are more readily masked than the lower frequencies, as discussed on p. 579, above, speech may be audible in a room with much background noise, and yet not intelligible.

CHAPTER 20

TEMPERATURE REGULATION

Heat is continually produced in the course of the various chemical reactions of metabolism. This heat must be lost to the surroundings and, except over short periods of time, the rate of loss must be equal to the rate of production and a state of *heat balance* maintained ; if this were not so, the body temperature would rise or fall indefinitely. The metabolic rate, and thus the rate of heat production, may vary over a wide range in different circumstances ; and a man may live in surroundings which are cold or hot, dry or wet, and lose heat readily or only with difficulty. He is able, nevertheless, to preserve heat balance with his temperature almost unchanged.

The processes by which the body temperature is prevented from rising or falling unduly are of two kinds. First, the rate at which heat is lost from the skin is increased or decreased as necessary (" physical " regulation). In man, this is done partly by wearing appropriate clothes, but as in all warm-blooded animals it is also done automatically. Like any other hot object, the body cools at a rate which depends on the difference between the temperature of its surface and the temperature of the air and other surroundings. It cools more rapidly also if it is wet than if it is dry, heat being absorbed by the evaporation of water, and the rate of cooling depends on the difference between the wetness of the body and the wetness (humidity) of the air. Physical regulation, accordingly, is exerted in two ways : (*a*) by raising or lowering the skin temperature through control of the *rate of blood flow through the skin* and thus the rate at which heat is brought out to the surface of the body ; and (*b*) by increasing the wetness of the skin, and thus the rate of cooling, by the *secretion of sweat*. Secondly, if the body temperature falls, in spite of all that can be done to reduce the loss of heat, the metabolic rate is increased, either deliberately by taking muscular exercise, or involuntarily by *shivering* (" chemical " regulation). The temperature of the skin, therefore, and at times that of a large part of the limbs, are subject to considerable variation. What is ordinarily referred to as the " body temperature," which remains nearly constant, is that of the " core," the internal organs of the abdomen, the heart, the brain, and the blood within these organs, although they are not all necessarily at exactly the same temperature. The temperature of a thermometer placed in the rectum, or under the tongue, is usually taken as a measure of the body temperature.

The Temperature Regulating Centres

The various processes by which a man keeps himself in heat balance are controlled and co-ordinated by " centres " in the brain in response

to signals transmitted from temperature receptors. Some of the signals must indicate the temperature of the "core," since it is this, primarily, which is regulated.

The results of changing the temperature of the blood flowing to the brain in the carotid arteries of an experimental animal are shown in Fig. 20. 1; the changes in the temperature of the tongue give a qualitative indication of those in the temperature of the brain. A rise in the brain temperature is accompanied by a fall in the rectal temperature; and a fall in the brain temperature by a rise in the rectal temperature. These experiments indicate that there are structures in the brain which are sensitive to changes in the temperature of the blood and are capable of so altering the rate of heat loss from the body as a whole that the

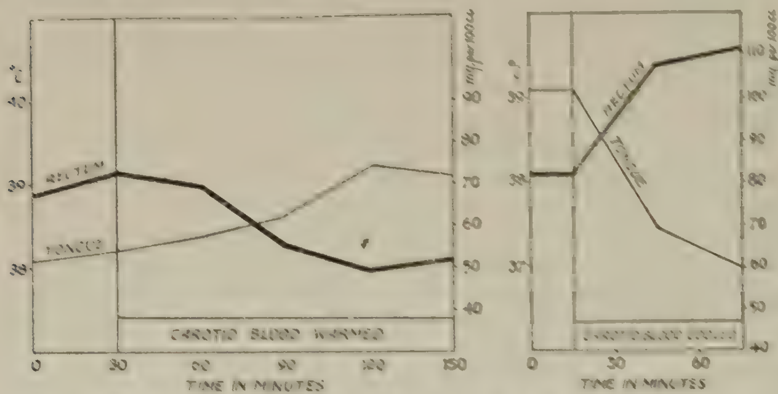


FIG. 20. 1. The Action of the **Temperature Regulating Centres.**

The effect on the body temperature of a dog, as measured in the rectum, of warming and cooling the blood flowing in the carotid arteries to the brain. The changes in temperature of the tongue are an indication of those in the brain. Note that the temperature of the rectum moves in the opposite direction to that of the brain, indicating the efforts of the regulating centres to counteract the warming and cooling of the blood supplying it. (From data by Geiger.)

changes in the blood temperature would be counteracted. Transection of the brain stem at various levels in experimental animals has shown that co-ordinated temperature regulation ceases if the section is made posterior to the hypothalamus (Fig. 14. 2, p. 391), so that the hypothalamus, and structures anterior to it, have no connection with the rest of the body: the "thalamic" animal has control of its temperature, but the "decerebrate animal" has little or none. More detailed studies, by localised electrical stimulation, and by warming and cooling small regions of the brain, show that the temperature receptors and the regulating "centres" lie close together in the anterior hypothalamus. It is probable that the organisation of the receptors and regulating centres is essentially similar in man to that in the experimental animals studied.

The regulating centres will initiate responses leading to increased

dissipation of heat, or to increased conservation or production of heat, only if they receive appropriate nerve impulses from temperature receptors. The body temperature cannot be regulated unless there is some change in temperature to initiate the regulation. The centres, however, receive nerve impulses from peripheral temperature receptors in the skin and the mucous membranes of the nose and mouth as well as from the central receptors. The temperature of the skin, as we shall see in more detail later, changes quite largely and rapidly in accordance with the temperature of the environment; the peripheral receptors, therefore, will " drive " the regulating centres appropriately even if there is no change in the temperature of the core. A man's body, moreover, has a considerable heat capacity, and appreciable amounts of heat must be gained or lost before there is a sufficient change in the temperature of the core to affect the regulating centres. Even if there were a sudden change in the rate of heat production, or in the temperature of the environment, the temperature of the central receptors would change only slowly, and the measures necessary to restore heat balance would be initiated gradually and after some delay. The peripheral receptors, however, will be affected more rapidly, and being very sensitive to changes in temperature, will immediately " drive " the regulating centres appropriately, thereby increasing the rapidity and precision with which heat balance is restored.

Normally, in a sedentary man, the temperature of the core varies from just below 36·5° C. (98° F.) to just below 37° C. (99° F.). Within these limits, there is a regular diurnal variation, the highest value being reached in the late afternoon, and the lowest in the early morning. In women, there may be a rhythmic variation with the menstrual cycle, the temperature being lower during the menstrual period and for some time afterwards. The body temperature rises and falls by about 1° C. (2° F.) in quite ordinary circumstances according to the temperature of the environment and the amount of clothing worn ; on exposure to severe cold it has been known to fall as low as 22·7° C. (73° F.) with subsequent recovery. It rises during muscular exercise in proportion to the severity of the exercise as measured, for example, by the oxygen consumption, and may reach 39·5° C. (103° F.).

When an infection is accompanied by a fever, it seems that the temperature regulating centres become " reset," as one might readjust a thermostat, to a higher value of the body temperature. One feels cold, one's skin is cold and pale, one shivers, and the temperature rises to the new value, often as high as 40° C. (104° F.) and possibly as high as 44° C. (112° F.). On recovery from the infection, one's skin becomes warm and flushed, one sweats, and the temperature is brought down again.

The temperature centres are depressed by anæsthetics and by hypoxia. In temperate climates anæsthetised animals and men must be kept warm by adequate coverings or an appropriate supply of heat. At high altitudes the air is cold and the partial pressure of oxygen is low ; breathing oxygen is of great assistance in preventing a fall in the general body temperature and in preventing frost-bite of the extremities.

The amount of control which can be exerted over the rates of heat

loss and of heat production, though large, are not unlimited. If the conditions are such that it is impossible to lose heat rapidly enough and the body temperature cannot be prevented from rising, the consequences are likely to be serious. The metabolic rate, like the rates of all chemical reactions, is increased by a rise in temperature. As the body temperature rises, the rate of production of heat will rise with it, and the temperature will rise even more rapidly. A vicious circle is established which will end in heat stroke and death. On the other hand, if in spite of all attempts to retain and produce heat the body temperature continues to fall, the temperature regulating centres, like other nervous structures, will become progressively less active. A time will come when the effort to maintain heat balance is abandoned ; the person " basks in the cold," and unless he is soon removed from the cold environment and warmed, he will die.

Control of the Body Temperature

All animals, including man, regulate their temperatures partly by means of appropriate behaviour. When warm, they keep quiet and seek shady and windy places ; when cold, they move about or seek sunny and sheltered places, curling up so as to reduce the amount of surface exposed. But we are here concerned with the less elaborate " reflex " processes which come more strictly within the scope of physiology.

Blood Flow through the Skin. An ordinary man feels comfortably warm when the temperature of his skin lies roughly between 25° C. and 35° C., although there are large individual variations. If he were at rest, and had no clothes, the air temperature would have to lie between about 20° C. (68° F.) and about 32° C. (90° F.), higher in dry windy conditions, when heat is carried away more rapidly, than in moist still conditions. In these " neutral " environments (neither hot nor cold) control of the body temperature is carried out entirely by control of the skin circulation. The excitation of the " 40° " receptors in the skin falls progressively as the skin temperature falls, and that of the " 30° " receptors rises reciprocally (Chapter 17, p. 515), whereas outside the " neutral " conditions, only one or other kind is excited.

If the rate of blood flow through the skin is large, heat is brought out rapidly from the core to the surface of the body, the surface temperature is well above that of the environment and the rate of heat dissipation is large. On the other hand, if no blood were to flow through the skin and superficial layers of the body, heat would reach the surface only by conduction through the layers of tissue, the core, in effect, having retreated away from the surface and become surrounded by an insulating layer ; the temperature of the skin would approach that of the environment and the rate of heat loss would be small. The flow of blood through the superficial layers cannot be stopped entirely, however, at least for any considerable period of time, since the tissues continue to use oxygen. But the cold blood returning to the heart from a cold skin flows in veins, the *venæ comites*, which form a network round

the arteries carrying the blood from the heart ; there is an exchange of heat, the venous blood being warmed and the arterial blood cooled (this has been observed experimentally). Heat is thus retained within the core even though the skin receives an adequate supply of blood. If the environment is warm, most of the blood flowing through the skin returns to the heart by way of the superficial veins and the "counter-current" heat exchange system is largely by-passed.

The magnitude of the vasomotor control over the vessels of the skin may be judged from the changes in skin temperature in different environments. The temperature of the bare skin is most accurately

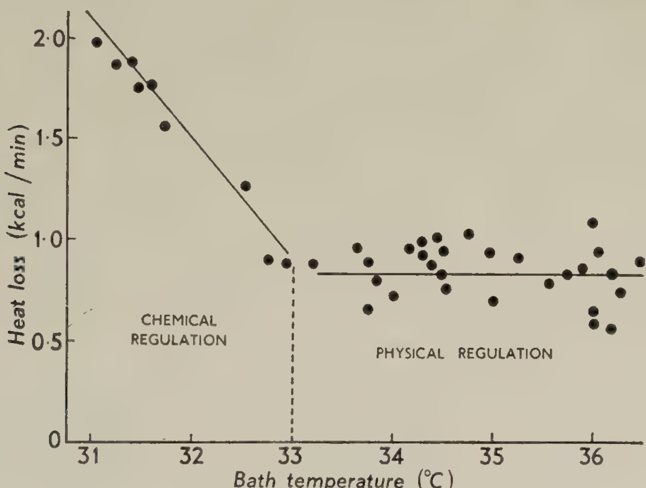


FIG. 20. 2. Temperature Regulation in a Water Bath.

A subject lay in a well-stirred water-bath, maintained at different temperatures on different occasions, and the rate of heat loss to the water measured : since his temperature was constant during the measurement, this must also have been his rate of heat production. It was constant at temperatures between 36·5° C. and 33° C., the range of "physical" regulation, by control of the cutaneous blood circulation. It rose steadily as the temperature was reduced from 33° C. to 31° C., "chemical" regulation, by increase in the metabolic rate, being added. (Burton and Bazett, from Burton and Edholm, "Man in a Cold Environment," Edward Arnold (Publishers) Ltd.)

measured in terms of the rate at which it radiates heat to an appropriately calibrated radiation thermopile held a few millimetres from the skin and protected from draughts. Alternatively, and rather less accurately though often more conveniently, thermo-couples may be attached to the skin by adhesive tape or thrust just below the surface of the skin ; this method must be used if the skin is covered by clothing. The greatest changes in skin temperature occur in the arms and legs, particularly in the fingers and toes, and the smallest on the head and the trunk. Different parts of the skin thus often have different temperatures, and the thermal insulation provided by the superficial layers varies from part to part. In cold conditions heat is lost more rapidly

from the head and trunk than from the arms and legs. When a man moves from warm surroundings, with a maximum flow of blood through the skin, to cold surroundings, with a minimum flow, the average thermal insulation round the whole core increases about five-fold. An average man can thus maintain heat balance, by control of his skin circulation in spite of a five-fold change in the difference between the body temperature and the air temperature, for example, or of a five-fold change in the rate of heat production by metabolism. (There are large variations between different individuals.)

The effectiveness of this control is illustrated in Fig. 20. 2. A subject was placed in a well-stirred water-bath, maintained at different temperatures on different occasions, and the rate at which heat was transferred from the subject to the bath was measured. At all temperatures between 33° C. (91·5° F.) and 36·5° C. (98° F.), this subject was able to keep his rate of heat loss almost unchanged, in spite of the fact that the difference between his body temperature and the temperature of the bath varied by a factor of nearly ten.

Local and Reflex Actions. These changes in the rate of blood flow through the skin are brought about : (*a*) by direct local action on the calibre of the blood vessels ; and (*b*) by reflex control of the calibre by varying excitation of the vasomotor nerve fibres supplying them, as discussed in Chapter 2 (pp. 65–67) and illustrated in Fig. 2. 17.

The immediate effect of exposing some part of the body, say a hand or foot, to a different environment, hotter or colder than before, is a vasodilatation or a vasoconstriction in the part exposed. The rate of blood flow through the hand, for example (almost entirely through the skin circulation) as measured by the rate of heat dissipation in a calorimeter (Chapter 2, p. 59), falls progressively as the temperature of the water is reduced, down to about 10° C. The consequent effect on the temperature of the skin, particularly on the tips of the fingers, when the hand is placed in cold water, is well shown in Fig. 20. 4 below. Again, the blood flow through the forearm, as measured by venous occlusion plethysmography (Chapter 2, p. 59), depends on the temperature of the water within the plethysmograph. If the temperature lies between about 15° C. and 20° C., the rate of blood flow is small, and little affected by temperature ; but if the temperature is greater than 25° C., the rate of blood flow increases rapidly with rise in temperature, by about five-fold at 35° C. and twenty-fold at 45° C. These measurements indicate that the blood flow through the muscles of the arm must be affected by temperature, as well as that through the skin.

A large part of the local vasomotor response, at least to cold, remains after removal, or blocking, of the sympathetic nerves and is produced either by an axon reflex or by a direct action on the blood vessels. The existence of *reflex* vasomotor control, through the temperature regulating centres, is shown by the response to change in the general body temperature or to change in the skin temperature of some other part of the body. In the experiment illustrated in Fig. 20. 3, the rate of heat dissipation from the subject's right hand was measured in a

calorimeter at a temperature of about 29° C. To begin with, his bare left forearm was in moderately warm air. On immersing this arm in water at a temperature of 15·5° C., there was an immediate vasoconstriction in the opposite hand, as shown by the decreased rate of heat dissipation to the calorimeter. This must have been due to a nervous reflex, set up by the temperature receptors in the cold arm, since the blood flow through this arm had been arrested before it was cooled, by inflating a sphygmomanometer cuff on the upper arm to 200 mm. Hg. The converse effect, of a reflex vasodilatation when some part of the body is warmed, may also be demonstrated in a similar manner, though not so regularly as the reflex vasoconstriction.

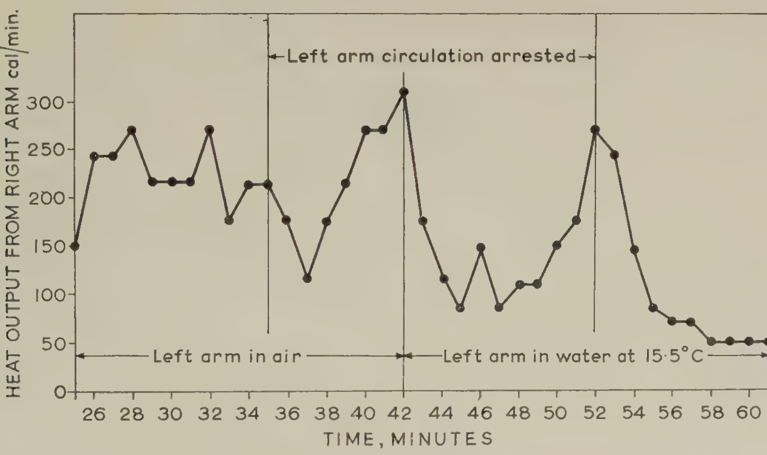


FIG. 20. 3. Reflex vaso-constriction in one hand, following cooling of the opposite arm.

The subject's right hand was placed in a water calorimeter at time 0, and the rate of heat elimination calculated from the rate of rise of temperature of the water, after 25 minutes had been allowed for a steady state to be reached.

Vaso-constriction is indicated by the reduction in the rate of heat output. It occurred : (*a*) at the 42nd minute when the left arm was placed in cold water, even though the circulation through this arm had been arrested and (*b*) at the 52nd minute when the circulation was restored and cool blood entered the general circulation.

(From Pickering, 1932, re-drawn. *Heart*, 16, 118.)

When blood was readmitted to the cold arm, and on returning to the general circulation cooled the central receptors, the blood vessels of the hand were again constricted, and remained so for a considerable time. In the absence of circulation through the cold arm, the heat loss from the body was not increased appreciably : there was no " drive " from the central receptors towards increased conservation of heat, the " drive " from the peripheral receptors was over-ridden and the vaso-constriction was transient, lasting only some seven minutes.

This interaction between signals from peripheral and central receptors is demonstrated, also, by the experiment illustrated in Fig. 20. 4. When

the subject held his hand in ice-cold water, stimulation of the peripheral temperature receptors led to an immediate constriction of the vessels of the nasal mucosa, as shown by the fall in temperature. The total effect of the reduction in heat loss produced by such vasoconstriction (there may also have been an increase in the metabolic rate) was apparently excessive: the rectal temperature rose slowly, and the vessels in the nasal mucosa dilated again. Later in the experiment the rectal temperature fell, and the mucosal vessels constricted more pro-

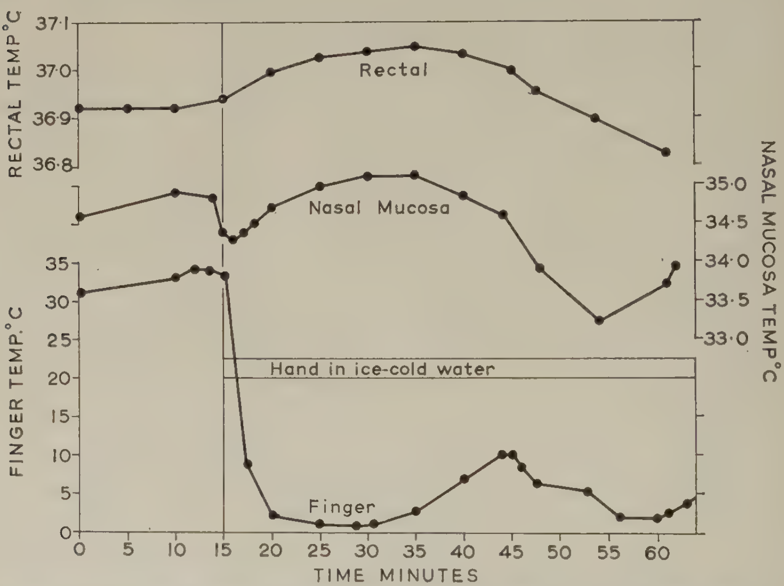


FIG. 20. 4. Temperature Regulation during extreme cooling of one hand.

The temperatures of the left index finger just beneath the skin, of the nasal mucosa, and within the rectum, are plotted against time. (Note the different scales on which the temperatures are plotted.)

After a control period of 15 minutes, the whole left hand was immersed in ice-cold water, producing extreme vasoconstriction in the finger, as shown by the great fall in temperature. After 20 minutes in the cold water, the temperature of the finger began to rise, owing to the onset of " cold vasodilatation ".

The fluctuations in the temperature of the nasal mucosa illustrate the vasomotor changes produced by interaction between the peripheral receptors in the cold hand and the central receptors responding to the rectal temperature. (From Aschoff, 1944, redrawn.)

foundly than before, the central and peripheral receptors now acting in conjunction.

When some part of the body, say a hand or foot, is placed in a very cold environment (air or water at a temperature below 10° C.), the tissues, particularly of the fingers and toes, may become so cold that there is risk of serious damage. This is avoided by the phenomenon known as "*cold vasodilatation*," produced by a direct local action on the blood vessels. Although protective locally, this increases the rate of heat

loss unless compensated by vasoconstriction elsewhere. As may be seen in Fig. 20. 4, after the subject's hand had been in ice-cold water for about twenty minutes, the skin temperature on the finger began to rise, indicating that the blood vessels were opening up again. The rectal temperature fell, and the blood vessels of the nasal mucosa became constricted in compensation. The vasoconstriction extended, also, after some delay, to the blood vessels of the cold hand, and its temperature fell again. Such a " hunting " process, of alternate vasoconstriction and vasodilatation, may continue so long as the hand (or foot) remains in the very cold environment.

The release of a profound vasoconstriction in the skin, either as a result of " cold vasodilatation " or more particularly of a sudden change to a warmer environment, is likely to lead to a rather sudden drop in the general blood temperature, owing to the initial return of cold blood from the skin. It is not uncommon for people to feel much colder, and even to start shivering, on first coming into a warm room after being out in the cold.

Reactions to Cold. Shivering. We have considered so far the effects of exposing only a small part of the body to cold environments. When a large part, or the whole, of the body is so exposed, the rate of heat loss may well be too great even when the blood flow through the skin has been reduced as far as possible everywhere. The first, and normal, reaction of an ordinary civilised man is then to put on more clothes. The effective thermal insulation of the stationary layer of air round the body can be increased, by this means, about six-fold, the limit being set by the fact that if the clothing is too thick one cannot move about : there is no such limit when a man is asleep. Fur-bearing animals have a comparable " pilomotor reflex," brought into action by the temperature regulating centres, by which the hairs are erected and the thickness of the air trapped between them is increased. When the thermal insulation provided by clothes or hair is sufficient, the temperature of the skin itself returns to the " neutral " zone, and further adjustment of the body temperature is made, as before, by control of the skin circulation.

If clothing is inadequate, the excessive heat loss is countered by increasing heat production. As may be seen in Fig. 20. 2, this occurred progressively, in the subject studied, when the temperature of the water in which he was immersed fell below 33° C. (91·5° F.). The subject's body temperature was constant when the measurements were made, so that the rate of heat loss to the bath must have been equal to the rate of heat production by metabolism. When the temperature of the bath had fallen to 31° C. (88° F.) the subject had increased his metabolic rate some two and a half times.

The greater part of the increase in metabolic rate results from increased muscular activity : in the absence of obvious muscular exercise, there is greater " tone " and rigidity in the muscles, and individual muscle fibres contract in an inco-ordinated and asynchronous manner, producing the irregular tremors of *shivering*. This peculiar

kind of muscular contraction is initiated by a " shivering centre " in the posterior hypothalamus ; this has direct connections with the temperature regulating centres and is set into action by a sufficient fall in the general body temperature. Shivering may also occur, however, when the skin becomes cold, or when very cold air is breathed, without detectable change in the rectal temperature. Most people, by this means, can increase the metabolic rate to about three times the resting rate. By deliberate muscular exercise, however, the metabolic rate may be increased ten-fold, or for a short time, more than a hundred-fold.

That part of the increase in metabolic rate which is not due to muscular contractions is thought to be due to increased secretion by the thyroid gland and the adrenal cortex, the temperature regulating centres being connected with the adenohypophysis and increasing the secretion of T.S.H. and of A.C.T.H. (Chapter 11, p. 337) ; the adrenal medulla also contributes by way of the sympathetic nerve supply. All these glands are known to secrete hormones which increase the metabolic rate of nearly all the organs and tissues of the body. They probably play an important part in the reactions to cold of some kinds of small mammal, but it is doubtful if they are of much importance in man.

Reactions to Heat. Sweating. The skin is always slightly wet as a result of the " insensible perspiration "; unless the air is saturated with water vapour, or the subject is immersed in water, there is always some loss of heat by the evaporation of water from the skin. The mucous membranes of the respiratory passages, also, are kept wet, and as pointed out in Chapter 4, the expired air is nearly saturated with water vapour at 37° C. Heat is lost by this route at a rate which increases with increase in the respiratory minute volume, and thus automatically with increase in the metabolic rate ; but in man it makes only a small contribution towards preservation of heat balance.

The wetness of the skin, and the loss of heat by evaporation of water, are greatly increased by the onset of active sweating, which occurs when there is a rise in the general body temperature as is shown in Fig. 20. 5. How large this rise must be depends on the skin temperature. In a cool environment, when the skin temperature is low, the rise in rectal temperature produced by a moderate amount of muscular exercise, for example, will bring about little or no sweating. In a warm environment, when the skin temperature is high, the same rise in rectal temperature will bring about copious sweating : if it is sufficiently warm, there may be sweating even without muscular exercise.

Sweat is produced by special glands in the skin which are innervated by fibres derived from the sympathetic system (Chapter 15, p. 456). They are brought into action by means of impulses in these fibres, derived from " centres " in the spinal cord, which may act to some extent independently of the hypothalamic centres : a spinal man may sweat when his skin is heated. The amount of sweat secreted in a given time by, say, an arm or leg may be measured by collecting it in a waterproof bag in which the limb is enclosed. It is not possible, however, to estimate in this way the rate of sweating by the whole body, since evaporation is prevented and heat balance upset. In studying tempera-

ture regulation, the rate at which water is evaporated is of greater interest than the rate at which sweat runs off the surface of the body. This may be measured by weighing the subject at the beginning and end of the period, and correcting for the volume of fluid drunk, the volume of urine eliminated, and the loss of weight consequent on the

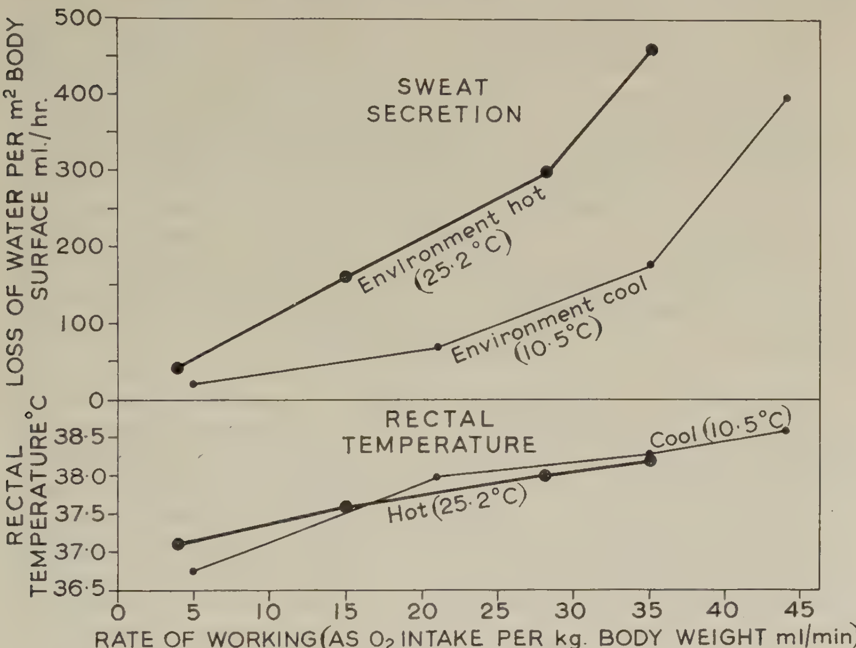


FIG. 20. 5. Secretion of Sweat during Exercise in Man.

Exercise was taken on a treadmill running at different speeds and with different gradients on different occasions. The severity of the exercise was measured in terms of the rate of oxygen consumption. Two series of experiments were performed in environments with effective temperatures (p. 602) of 10·5° C. (51° F.) and 25·2° C. (77·5° F.) corresponding roughly in England to a mild winter day and a hot summer day respectively.

The rate of sweat secretion (measured as the rate of evaporation of water and including insensible perspiration) increased rapidly with increase in the metabolic rate, and was much greater in the hot environment (skin temperature about 34° C.) than in the cool environment (skin temperature about 28·5° C.). The skin temperature did not rise when exercise was taken.

The rectal temperature increased progressively with the severity of the exercise ; it varied with the environmental temperature when the subject was at rest, but hardly at all when he was taking exercise. (Replotted from data of Robinson *et al.* in Newburgh " The Physiology of Heat Regulation," W. B. Saunders Co.)

loss in the expired air of the carbon of the foodstuffs metabolised—*i.e.* the difference between the weight of oxygen absorbed and the weight of carbon dioxide eliminated : this may be calculated from the metabolic rate. In extreme conditions a man may secrete 10–15 litres of sweat a day and, for a short time, evaporate water at a rate of 1 litre an hour.

The sweat glands secrete a vasodilator substance along with the sweat (Chapter 2, p. 67), so that there is a plentiful supply of blood, and thus of heat to evaporate the sweat. In parts of the skin which are particularly well supplied with sweat glands, this active vasodilatation largely replaces the passive vasodilatation, following inhibition of the vasoconstrictor nerves, which occurs in other parts of the skin.

There are two kinds of sweat gland, known as *eccrine* and *apocrine*. The apocrine glands, which occur chiefly in the axillæ and pubic regions, are not concerned in temperature regulation, but probably have some secondary sexual function. The eccrine glands may be brought into action, not only by thermal stimuli, but also by emotional and mental stimuli, as in the " cold sweat " of fear and anxiety : the glands in the palms of the hands and soles of the feet, indeed, appear to respond only to such stimuli.

Some kinds of animal (dogs, for example) are not provided with sweat glands. They make use of the evaporation of water for dissipating heat by means of *panting*. When the body is hot, the respiration becomes very fast and shallow, so that a large amount of air passes over the surface of the tongue and outer respiratory passages. The amount of air which enters, and ventilates, the lungs is not increased, since there is no change in the composition of the alveolar air even after several hours of panting.

The rate at which heat is lost by evaporation of sweat depends not only on the rate of secretion, but also on the climatic conditions and on the access of air to the skin as affected, for example, by winds and clothing. In hot damp conditions the rate of secretion of sweat may exceed the rate of evaporation, much of the sweat may run off the skin and adequate cooling is difficult. When the atmosphere is dry, on the other hand, the rate of cooling may be large enough to counteract an actual gain of heat from an environment whose temperature is above that of the body. Indeed, a man may keep his temperature constant in an atmosphere with a temperature as high as 120° C. (250° F.), sufficient to cook meat or to boil an egg, provided that the air is dry.

Sweat is not pure water, but is a solution containing, on the average, about 0·3 per cent. sodium chloride—about one-half the concentration in plasma ; there are very large variations in different circumstances, between different individuals, and between different parts of the same individual. It contains, also, most of the other constituents of the plasma except the proteins ; lactate ions and urea, indeed, are usually present in greater concentration than in the plasma. Sweat which runs off the skin without evaporating is not entirely useless, since it washes away the dissolved substances which would otherwise be left on the skin, increasing the concentration of the sweat and reducing the rate of evaporation.

The loss of water by sweating results in the sensation of thirst, and the loss is made good by drinking (Chapter 7, p. 223). The loss of salt, however, may have important consequences. Failure to take in enough salt when recovering from heavy sweating will result in a dilution of the body fluids ; ordinarily the excess water will be rapidly eliminated by the kidneys (Chapter 9, p. 272) and the failure to restore the water balance may produce some discomfort but nothing more. If, however,

the kidneys are for any reason unable to eliminate the excess water, a more serious condition of *water intoxication* may be produced. Even in the absence of renal insufficiency, this may occur when severe exercise is taken in hot surroundings and pure water is drunk while the exercise continues. Severe exercise largely inhibits water diuresis, partly owing to diversion of the blood from the kidneys to the active muscles and partly by an action on the neurohypophysis leading to a secretion of ADH. The consequent dilution of the body fluids then gives rise to the incapacitating condition known as " miners' (or stokers') cramp." (The condition was first observed in miners and in stokers of coal-fired boilers.) The necessity for taking an adequate quantity of salt along with the water (as occurs naturally in beer, for example) was in fact known empirically among miners and stokers before its *rationale* was understood.

The loss of substances other than salt in sweat is ordinarily of no consequence, since they would otherwise be excreted by the kidneys. If the kidneys are severely diseased, however, elimination of substances like urea by copious sweating may be of substantial benefit.

The salt concentration of the sweat depends on the state of the salt balance of the whole body. Prolonged existence in hot surroundings, with more or less continuous sweating, is apt to lead to some degree of chronic salt depletion ; this is believed to account for the observed decrease in the salt concentration of the sweat as one becomes " acclimatised " to the environment. It does not occur if enough salt is taken to preserve the balance. The rate at which salt is absorbed from the sweat back into the blood stream, like the rate of sodium reabsorption in the renal tubules, is regulated by means of the hormone *aldosterone*, elaborated by the adrenal cortex ; the rate of secretion rises as the sodium content of the body falls (Chapter 11, p. 326).

Climate and its Measurement

It is often useful to be able to measure the properties of an environment which affect the ease with which people living and working in it can regulate their body temperatures. The comfort and efficiency of workers in factories and offices may be affected by the atmosphere in which they have to work ; the nature and quantity of clothing, food and drink to be taken on expeditions to polar, mountainous or desert regions are decided by the nature of the climate to be expected.

The *temperature* of the air, other things being equal, determines the rate at which heat is lost from the body by conduction and convection ; the *humidity* of the air determines the rate at which heat is lost by the evaporation of water. Both these quantities may easily be measured by means of dry and wet bulb thermometers, sheltered from the sun and the wind.

In given conditions of temperature and humidity, the rate of heat loss is greatly affected by the *wind speed*, which can be measured by instruments known as anemometers. The air in immediate contact with the skin (or the surface of any hot object) forms a " stationary layer," usually a few millimetres thick ; beyond this, the air is in motion relative to the skin as a result of thermal currents (hot air is

less dense than cold air, and so rises), movements of the person whose temperature is being considered, and above all, the presence of winds. Heat and water vapour must pass through the stationary layer before being carried away by the moving air : the greater the air movement, the thinner is the stationary layer and the more rapid is the passage of heat and water vapour ; hair and clothing, conversely, increase the thickness of the stationary layer and decrease the rate of heat loss. In the absence of appreciable evaporation of water, and for a given difference in temperature between the skin and the air, the rate of heat loss in a 10 m.p.h. wind is about four times as great as it is in still air, and in a 50 m.p.h. wind about nine times as great. For this reason, the unpleasantness of being in warm moist air may be largely removed by stirring the air with a fan.

Lastly, heat may be lost, or not infrequently gained, by radiation to or from surrounding objects which absorb or emit heat, to interstellar space by night or from the sun by day. The surrounding objects will, in general, be at different temperatures, have different shapes and have surfaces of different radiating efficiencies ; measurement of their effect on the heat balance of a man is not easy, and is necessarily somewhat empirical. The " solar heat load," however, can be deduced fairly accurately from the altitude of the sun and the amount of mist or smoke in the atmosphere.

In exceptional circumstances—at least in the ordinary life of civilised man—substantial amounts of heat may be lost to rain or snow which falls on the skin or clothing. If the clothing (or an animal's hair) becomes wet, moreover, the trapped air is replaced by water and the thermal insulation is greatly reduced. Wet clothing is better than nothing, however, since the water within it is trapped and delays the conduction of heat to the rain or snow which falls on the outer surface.

Various attempts have been made—without marked success—to combine all the relevant properties of the climate into one figure which defines its overall cooling (or heating) power. From such figures, however, it is possible to make useful estimates, for example, of the amount of clothing needed in a given climate by a man taking different amounts of exercise. An empirical figure called the " effective temperature " is defined as the temperature of a still atmosphere, saturated with water vapour, which feels, subjectively to an ordinary person, the same as the particular climate under investigation. It may be deduced from measurements of the temperature and humidity of the air and the wind speed (and also the effects of radiation for the " corrected effective temperature "), by means of charts constructed from empirical observations on human subjects.

Alternatively, one may use the readings of the *katathermometer*, suspended in the place where observations are to be made. This is an alcohol thermometer with a large bulb ; the cooling power of the air is measured as the reciprocal of the time taken for its temperature to fall from 100° F. to 95° F. When dry, the katathermometer measures the rate of heat loss by conduction, convection and radiation ; when

the bulb is covered by a moist finger-stall, it measures also the heat loss by evaporation of water. The surface of a man is rarely completely wetted, and the cooler and drier is the air, the smaller is the fraction wetted. In ordinary circumstances, this fraction lies between 10 per cent. and 50 per cent., and an appropriate figure between the two values obtained with the katathermometer should be taken as a measure of the cooling power of the climate.

CHAPTER 21

REACTIONS TO INJURY

In order to keep alive we must be in close contact with the outside world—with the air, water, food and all kinds of solid objects which it contains. It is a matter of common knowledge that in doing so we may suffer injury. Our skins and the mucous membranes of our noses, mouths and throats may be damaged by cuts and grazes, or by excessively hot or cold materials ; but for the most part, these are of little consequence. Even if the injury is sufficiently deep to open a blood vessel, little real harm is done provided that the shed blood clots properly ; if it does not, the loss of blood is inconvenient, to say the least, and may become so great as to be fatal. Clotting, or coagulation, of the blood, then, is the first of the reactions to injury which we shall consider.

The surrounding world, also, contains large numbers of micro-organisms, protozoa and bacteria, and viruses. These may enter the body through the skin and the mucous membranes, particularly if damaged ; some of them become parasites, multiply in the body fluids and cause disease. The various processes by which such an invasion is opposed form the other group of reactions to injury which we shall consider. These reactions are brought into play, also, by the presence in the body fluids of many different kinds of " foreign " material, many of which are not micro-organisms, or any kind of living organism, and not necessarily injurious. As a result of studying these reactions, we have learnt that each individual is not only unique in its physical structure and appearance, but also in its chemical composition : cells of one individual, if transferred to another, are treated as " foreign," potentially injurious, and destroyed. One cannot, in general, replace organs or tissues lost by one individual by those taken from another individual, although there are a few exceptions. Fortunately, the red blood cells are among these exceptions, and are less " unique " than are those of other kinds of cell ; those of any individual may be placed into one or other of a small number of groups. The blood of one individual may be " compatible " with that of another, so that blood lost by one may be replaced by that of the other.

The Coagulation of the Blood

When a blood vessel is opened and the blood flows over the surrounding tissues, it normally sets to a jelly in the course of five or ten minutes. It is then said to have *clotted* or *coagulated*. If the clot is collected and washed free from red cells, it is found to consist of an interlacing network of fine white fibres of protein material to which the name *fibrin* is given. If a blood clot is allowed to stand, it slowly shrinks (retraction or syneræsis) and a pale yellow liquid called *serum* is squeezed out. Serum

closely resembles plasma, but lacks a protein constituent which can be precipitated from plasma by half-saturation with sodium chloride. This protein is called *fibrinogen*, because it changes into fibrin during the process of clotting. If precipitated fibrinogen is redissolved in 2 per cent. sodium chloride, and a little fresh serum added, a clot is formed in much the same way as it is in plasma. Fresh serum therefore contains a substance which induces fibrinogen to clot ; this substance is known as *thrombin.*

Thrombin is an enzyme, and may be isolated in a moderately pure state by extracting fresh serum with alcohol, or by various other more elaborate procedures. Unclotted blood treated in the same way yields no thrombin, so that thrombin does not exist in circulating plasma, but is formed some time after the blood is shed. Plasma, therefore, must contain a substance—known as *prothrombin*—which can change into thrombin under the proper conditions. Prothrombin is ordinarily formed continuously in the liver, an adequate supply of a special vitamin (vitamin K—see p. 216) being necessary. If the supply is inadequate (owing to an inadequate diet, for example) there is partial failure of blood clotting.

There is a deficiency of vitamin K, also, when, owing to disease, the bile duct is obstructed ; it is then not properly absorbed from the alimentary canal. The formation of prothrombin is impaired, as might be expected, when the liver is damaged as a result of disease or poisons such as chloroform. In all these types of inadequate liver function, however, the failure of blood clotting does not result only, or even chiefly, from an inadequate supply of prothrombin, but also from an inadequate supply of factors V and VII (see below, p. 609).

A failure of blood clotting, generally similar to that produced by inadequate liver function, is produced by administration of a substance known as *dicoumarol.* This is a derivative of the substance coumarin, which is responsible for the odour of new-mown hay ; dicoumarol may be formed from it if sweet clover is improperly cured before being stacked. Animals eating this hay then develop a hæmorrhagic disease.

Since blood does not normally clot in the circulation, we infer that the conditions necessary for the conversion of prothrombin into thrombin, and the initiation of the process of clotting, are : (1) the contamination with tissue juices which takes place as blood flows from a wound, and (2) the contact with foreign surfaces such as the skin and the vessels in which the blood is collected. In these conditions, it appears that a substance called *thromboplastin* is released. This is most probably a second enzyme, and is present in extracts of most tissues, since the addition of such extracts to blood increases the rate of clotting. But it is derived particularly from very small cellular elements normally present in the blood, called *platelets* or *thrombocytes.* These are round or oval disc-shaped bodies, 2 to 3μ in diameter, with granular cytoplasm, but no nucleus. They are believed to be derived from large multinuclear cells of the bone marrow called megakaryocytes. Cytoplasmic processes of these cells become pinched off and pass into the blood as platelets. Normal blood contains from 250,000 to 450,000 platelets per cu. mm.

Within a few seconds after blood is shed its platelets agglutinate, that is, clump together, and then, more slowly, disintegrate. Consequently, in smears of normal blood the platelets are to be seen only in clumps often consisting of as many as thirty or forty platelets. The clumping of platelets at the site of bleeding in small vessels tends to plug the wound in the vessel and thus helps to restrain hæmorrhage ; their subsequent disintegration releases thromboplastin, and so hastens blood clotting. If the early stages of clotting are observed by means of an ultra-microscope, needles of fibrin are seen to form in the immediate neighbourhood of clumps of disintegrating platelets, and to grow out from these centres until the whole of the blood is enmeshed in the network.

Agglutination and disintegration of platelets proceeds rapidly when blood is in contact with some kinds of " foreign " surface—such as glass and other substances which are wetted by blood—and only slowly if blood is shed through a cannula into a receptacle, both made of a suitable plastic such as polythene, or coated with silicone or paraffin wax ; these surfaces are chemically inert and are not wetted by the blood. If the platelets disintegrate rapidly, clotting is rapid ; if they remain intact for some time, clotting is delayed, but can be brought on rapidly if disintegrated platelets are added.

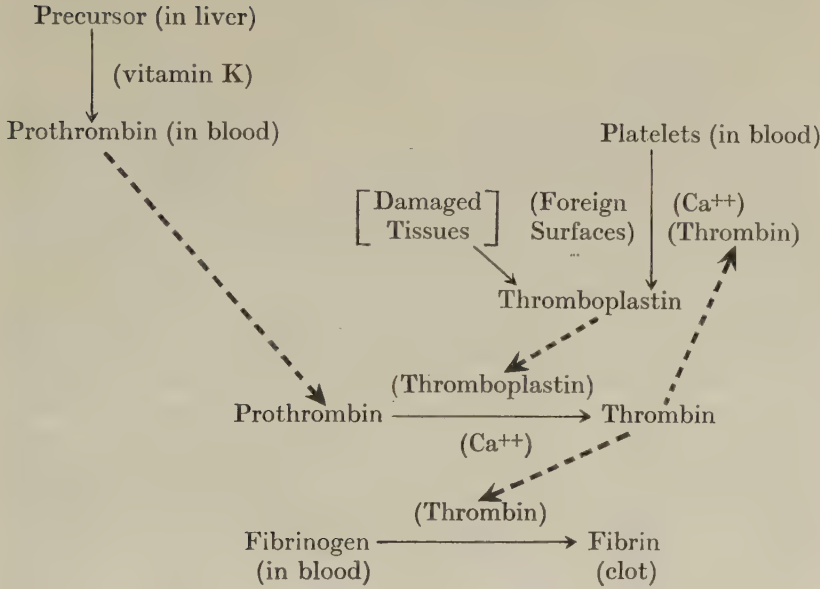
A further condition for the conversion of prothrombin to thrombin, and for the release of thromboplastin from the platelets, is the presence of calcium ions : if these are removed from freshly shed blood, by precipitation as calcium oxalate, by combination in an un-ionised state as calcium citrate, or by " chelation " with ethylene diamine tetracetate (EDTA) clotting does not occur.

It is probable that prothrombin is not converted into thrombin directly by the substance called " thromboplastin " (also known as *thrombokinase*) ; and the thromboplastin of tissue extracts is probably not identical with that of the platelets (at one time called *cytozyme*). The name " prothrombinase " has been given to the substance derived from tissue extracts which is responsible for the formation of thrombin ; it has not been established that this same substance appears in the platelets in the absence of tissue extracts. This, however, does not affect the essential nature of the sequence of reactions which brings about clotting.

The process of blood clotting, then, may be regarded as resulting primarily from the interaction of three substances normally present in the blood—fibrinogen, prothrombin and calcium ions—with a fourth substance—thromboplastin—released from the platelets as a result of their disintegration. This release, moreover, is accelerated by the presence of thrombin, which thus indirectly accelerates its own formation from prothrombin ; the reaction sequence is autocatalytic, and once started, proceeds at a progressively increasing rate. This accounts for the fact that even in the presence of all the factors necessary for clotting, there is a delay of several minutes before any change occurs, and then there is an almost explosive appearance of the clot. The whole reaction sequence is summarised, schematically, in Table 21. 1.

TABLE 21.1

The Clotting of the Blood



The fact that the circulating blood does not clot spontaneously and continuously may be very reasonably attributed to the stability of the platelets, and the consequent failure of prothrombin to be converted into thrombin. But there is good reason to believe that there is a second line of defence in the presence of naturally occurring anti-coagulant substances, the most important of which is *heparin*. This acts both in preventing the formation of thrombin from prothrombin, and in antagonising the action of thrombin on fibrinogen. Heparin was originally isolated from the liver (hence the name), but was later shown to be contained particularly in the mast cells of the reticulo-endothelial system (see p. 611, below). These cells are present in considerable numbers in the connective tissue which surrounds the small blood vessels, and it is possible that they release heparin into the blood stream.

Shed blood can be preserved in the fluid state in the following ways :—

(1) By defibrination. The blood while clotting is stirred with some object, *e.g.*, a bundle of feathers—to which fibrin will adhere. The fibrin is thus removed as it forms and the red cells are left suspended in serum.

(2) By precipitating the calcium by the addition of sodium or potassium oxalate (0·1 to 0·3 per cent.).

(3) By removing calcium ions by the addition of sodium citrate (0·2 to 0·4 per cent.), fluoride or EDTA. Since calcium ions are necessary not only for the change of prothrombin to thrombin, but for the disintegration of platelets, oxalate and citrate solutions can be used to preserve platelets intact.

(4) By cooling the blood to 0° C., which retards clotting almost indefinitely and preserves the platelets.

(5) By preserving blood from contact with surfaces which it wets.

(6) By the addition of heparin.

(7) By the addition of certain azo dyes, *e.g.*, chlorazol fast pink, which are thought to act partly as antithromboplastins.

(8) By the addition of hirudin, a material obtained from leech heads which acts as an antithrombin.

(9) By the addition of suitable concentrations of almost any neutral salt, *e.g.*, one-seventh saturation with $MgSO_4$.

Blood clotting is of great value when localised near bleeding points due to injury or surgery. If, however, clots become detached from such points, or arise spontaneously, and travel as emboli in the blood circulation, they become a great danger. Often they lodge in the lungs (pulmonary embolism). Moreover, clots may form in diseased blood vessels and by blocking them, may impair vital functions as in coronary thrombosis or cerebral thrombosis. In embolism or thrombosis, dicoumarol or more commonly synthetic anticoagulants are administered. When blood is passed through external apparatus and back into the body, as in many kinds of physiological experiment on animals, or when an " artificial heart " is used to maintain the patient's circulation during operations on the heart, clotting is likely to be initiated by the materials used in the apparatus. In these circumstances, heparin is injected to prevent the undesirable clotting. Blood which is withdrawn from a donor, for subsequent transfusion to a patient who needs it, is ordinarily prevented from clotting by the use of citrate, heparin being relatively expensive. Citrate, however, cannot be used to prevent clotting in the whole of the blood in the circulation, since the complete removal of calcium ions would have disastrous effects on a great many of the bodily functions (compare Chapter 7, p. 212 and Chapter 11, p. 332). It may be desirable, on the other hand, to promote clotting in order to check hæmorrhage when blood vessels have been opened during surgery or as a result of accident ; it is usually sufficient to provide a large area of " foreign " surface by applying a swab of cotton (the traditional cobweb is useful for the same reason, but is less desirable owing to the risk of infection), but in severe cases the swab may be soaked in a solution of thrombin.

The problem of checking hæmorrhage may become really serious in certain persons, known as " bleeders," who suffer from an inborn disease in which blood clotting is so slow that hæmorrhage from a relatively trivial wound may be profuse enough to be dangerous : this *hæmophilia* is an hereditary defect which is manifested almost exclusively in males, but inherited through the mother. A person suffering from this disease—or more strictly, group of diseases—is found to lack none of the primary components of the clotting reactions (as summarised in Table 21. 1). Detailed study of these reactions, however, and progressive purification of the known components, both in normal blood and in the bloods of hæmophilic subjects, has revealed the existence of several additional " factors " which must be presumed to take part in the formation of thromboplastin and in the conversion of prothrombin to thrombin. These " factors " appear, ordinarily, to be associated with

the globulin fractions of the plasma proteins, and it is not easy to demonstrate their presence. The hæmophilias result from genetic deficiencies, such that one or other of these factors is not synthesised, and the clotting reactions do not proceed as they should.

Up to 1958 at least six of these additional factors had been described, and there may be more, since the nomenclature has been extremely confused, and the problem is complicated by the presence of anti-coagulants and inhibitors which oppose their actions. They have been identified by means of the Roman numerals from V upwards, with the exception of VI (given originally to a factor which was later found not to have a separate existence) ; the first four—and the most important—factors are those substances already recognised as essential for blood clotting, and given in Table 21. 1. Many names have also been given to them describing their origins and supposed functions. Factor V is labile, disappears from oxalated plasma on storage in a refrigerator, and is consumed when fresh plasma clots. Factor VII, on the other hand, appears or becomes activated, when plasma clots, and thus accentuates the autocatalytic nature of the whole process. Factors V and VII were discovered in normal blood by means of these and other specific properties : they are necessary, in addition to tissue extracts and calcium ions, for the formation of thrombin from prothrombin.

Other factors have been discovered by study of the hæmophilias, the commonest of which is known as " true hæmophilia " or " hæmophilia A," and others are known by the names of the patients who were first found to be deficient in factors not previously recognised. To give two examples only, it has been found that factor VIII is lacking in patients with " true " hæmophilia, and factor IX is lacking in patients with a type of hæmophilia first observed in a man with the name of Christmas (it is thus called the " Christmas factor "). Factor V and factors VIII and IX must be present, in addition to calcium ions, for the formation of thromboplastin from the platelets.

Blood clotting may be defective, also, owing to a reduction in the number of platelets in the blood, or, very rarely, owing to a lack of fibrinogen. These deficiencies may be congenital, or may result from disease.

Measurements of Clotting Time. The clotting time of normal blood may vary from four to sixty minutes, depending on the method used to measure it. To get consistent results, the following conditions must be controlled : (1) *Temperature.* Clotting time increases as the temperature decreases. (2) *The manner of obtaining the blood.* Blood drawn from a vein clots more slowly than blood from a skin puncture, which allows more contamination by tissue juices. (3) *Agitation of the blood.* Agitation hastens clotting. (4) *Cleanliness of apparatus.* The cleaner the apparatus, the slower the clotting.

A simple method is to collect a few drops of blood from a puncture in the lobe of the ear on a clean watch glass, which is then covered by another watch glass to limit evaporation. The fluidity of the drop is tested from time to time by gently tipping the watch glass. The time from the shedding of the blood till the first signs of clotting appear under these conditions is about four to eight minutes at 20° C., for normal human blood.

Detection of abnormalities in the clotting reactions starts by the use of such simple procedures. But if clotting is found to be abnormally slow, more elaborate procedures must be used in order to establish the nature of the defect.

Reactions to " Foreign " Materials in the Body Fluids

When " foreign " or abnormal substances or living organisms enter the body, they are removed, and any deleterious actions produced are opposed, by a number of processes which are often referred to collectively

as "defence reactions." Particulate matter, whether living or dead, is removed by absorption (*phagocytosis*) in certain kinds of *leucocyte*, which circulate in the blood, and in certain tissue cells belonging to the *reticulo-endothelial system*. Certain substances, usually but not necessarily associated with, or derived from, "foreign" organisms, react in various ways with substances already present in the body fluids of the animal or person invaded (or specially formed in response to the invasion), the reactions being of an unusual nature and called "immunity reactions"; they were first studied in connection with the fact that after a person has recovered from certain kinds of disease, he may subsequently be unharmed by—or immune to—further attacks by the particular micro-organism which causes the disease.

The Leucocytes or White Blood Corpuscles. These are usually classified into three groups: the granular cells (sometimes called *Polymorphs*, but now often called *Granulocytes*), the *Lymphocytes* and the *Monocytes*.

The *Granulocytes* are called "polymorphs" (polymorphonuclear leucocytes) because in stained smears of blood their nuclei are seen to be divided into two, three or four lobes by deep indentations. They are sub-divided, according to the staining reactions of their granules, into:

Neutrophils, with very fine cytoplasmic granules, showing no striking affinity for acidic or basic stains (the name "polymorph" is often restricted to this group of granulocyte);

Eosinophils, with very coarse granules, stained bright red by eosin, the nucleus being nearly always bi-lobed; and

Basophils, with coarse granules having an affinity for basic dyes and hence stained blue by the stains commonly used for blood.

The red bone marrow is the normal site of production of granulocytes and in it can be recognised immature granular cells in two developmental stages. The younger are known as myeloblasts and the older as myelocytes. Myeloblasts have no granules, but myelocytes have granules in which the tendency to become eosinophilic, basophilic or neutrophilic can be easily seen. The nuclei of the myelocytes, however, are only slightly indented. These immature forms of granular blood cells may appear in the circulating blood in blood diseases, called *leukœmias*.

The *Lymphocytes* of normal blood are the smallest of the white blood cells, the small forms having about the same diameter as a red cell. They are characterised by an almost round, densely staining nucleus surrounded by a narrow zone of cytoplasm free from granules. The lymphocytes are produced in the lymphoid tissue from larger cells with pale nuclei.

The *Monocytes* (*transitional cells* or *large mononuclears*) have large oval or bean-shaped nuclei; both nuclei and cytoplasm are pale staining.

Normal human blood contains 5,000 to 10,000 leucocytes per cu. mm. When the number is lower than normal, a state of *leucopenia* is said to exist—as, for example, in typhoid and influenza. Most general infec-

tions, on the other hand, are accompanied by a striking increase in the number of circulating leucocytes, a condition known as *leucocytosis*.

The number of leucocytes per cubic millimetre of blood (the *White Cell Count*) is estimated in the *hæmocytometer* in much the same way as is that of the red blood cells (see Chapter 3, p. 70). The blood is diluted 20-fold, in a special pipette, with a fluid which contains acetic acid to dissolve the red cells, and suitable stains for the white cells. A drop of the suspension so formed is placed on the special microscope slide, forming a layer exactly 1/10 mm. deep. When the cells have settled, the number lying within a square 1/16 sq. mm. in area is counted. These squares are formed by lines ruled at the corners of the area containing the much smaller squares used for the red cell count. Several such counts are made, and the average count is multiplied by 160 (the cells counted came from 1/160 cu. mm. of suspension) and by 20 (the dilution factor) to give the number of cells per cubic millimetre of the original blood.

The estimation of the proportion of each type of leucocyte in a particular sample of blood can be done on any properly stained blood smear, and is known as the *Differential White Count*, approximate values for normal human blood being :

Granulocytes	Neutrophils,	70 per cent.	Of total white cell count.
	Eosinophils,	1 ,, ,,	
	Basophils,	0·5 ,, ,,	
Lymphocytes		24 ,, ,,	
Monocytes		4 ,, ,,	

The leucocytes are motile, and can creep over surfaces and make their way out of the blood into the tissue fluids through interstices in the blood vessels. The lymphocytes, in particular, circulate from the blood stream into the tissue spaces, whence they are carried to the lymph nodes with the lymph, and then back again into the blood stream by way of the thoracic duct.

The Reticulo-endothelial System. The loose, or reticular, connective tissue which surrounds the blood vessels, for example, and provides support and attachments for the abdominal organs, contains, besides the fibroblasts which are responsible for the reticulated structure, other kinds of cell called mast cells, plasma cells and macrophages. The same, or very similar, kinds of cell are found in the endothelia lining the sinusoids of the liver, spleen and bone marrow, and in the lymph glands. These cells, in their various locations, constitute the "reticulo-endothelial system," and they are closely related to the leucocytes of the blood. The whole subject is very confused, but it is thought that the basophil leucocytes may well be identical with the mast cells of the reticulo-endothelial system ; and lymphocytes seem, in certain conditions, to develop into macrophages, with monocytes perhaps as intermediate stages.

Phagocytosis. The most important of the cells which ingest bacteria and other foreign particles, living or dead, are the macrophages of the reticulo-endothelial system. When dyes, such as trypan blue or lithium carmine, or particulate matter such as indian ink, are injected intra-

venously, they are taken up for the most part by these macrophages, particularly by those in the liver known as Kupffer cells. When there has been a local invasion of bacteria, as in an infected wound of the skin, there is a great accumulation of neutrophil granulocytes (polymorphs), which leave the blood stream in the infected region, ingest the bacteria, and release proteolytic enzymes—perhaps as a result of their own disintegration ; these enzymes break down cells which have been killed by the infection. The liquid known as pus is made up of these broken down cells, together with the remains of the leucocytes, and possibly some erythrocytes : it is sealed off from the body fluids in general by the proliferation of the fibroblasts, which form the scar tissue.

Inflammation due to local infection is accompanied by dilatation of the blood vessels in the region affected. This results largely from the action of histamine, or histamine-like substances, as in the "triple response" produced by mechanical or chemical injury to the skin (Chapter 2, p. 68). The dilatation of the blood vessels hastens the accumulation of leucocytes in the region and facilitates their escape into the tissue spaces. The mast cells of the reticulo-endothelial system, and the basophil leucocytes, contain a precursor of histamine, and may well be the main source of the histamine released in tissues damaged by wounds or by bacterial "toxins."

Immunity Reactions. Certain foreign substances, many or all of which are proteins, when introduced into the body fluids, act as *antigens*, and lead to the development of *antibodies*, which react with them in various ways. If the antigen is a protein in solution—such as a serum protein of a different kind of animal—the antibody which is formed precipitates the antigen, a reaction which can be easily observed in a test-tube ; the antibody is thus termed a *precipitin*. Antigens, however, may also be present on, or released by, living cells such as bacteria or red blood cells. The corresponding antibody will then : (*a*) make these cells stick together in clumps, or agglutinate, when it is called an *agglutinin* ; or (*b*) lead to their destruction or lysis, when it is called a *lysin* (or more specifically, for example, a *bacteriolysin* or a *hæmolysin*) ; or (*c*) neutralise the poisonous (toxic) effects of the substances released by bacteria, when it is called an *antitoxin*. Antibodies of all kinds seem to be produced chiefly by lymphocytes ; but at the site of an infection there is a great accumulation of eosinophil granulocytes, and these also may play some part in the antigen-antibody reactions.

The antigen-antibody reactions are highly specific, most kinds of antibody reacting only with one particular kind of antigen ; serum proteins from different species, which cannot be distinguished by chemical methods, can be distinguished by the way they react with an antibody prepared against one of them.

It is the presence of the antibodies formed in response to an invasion of bacteria, or other micro-organisms, which is responsible for the subsequent failure of further infection to produce the disease—an immunity which may last for many years. Many of these antibodies

can be prepared artificially, or the animal or man can be induced to develop them itself without having to undergo the full rigours of an actual attack of the disease ; their study has consequently become an important part of medical science, known as *Immunology*.

Complement. Suppose a specific hæmolysin has been produced in one animal by repeated injections of the red cells of another animal ; if serum containing the specific hæmolysin is heated to 56° C. its hæmolytic power is lost. The addition of fresh serum from almost any normal animal, however, restores the hæmolytic power. Since the added normal serum would not, by itself, have hæmolysed red cells, it is evident that two factors are concerned in this type of hæmolysis, (1) a specific antibody which is stable to heat (thermostable), and (2) a non-specific factor which is destroyed by heat (thermolabile), known as the *complement* ; this is present in any normal serum. Red cells which have been treated with *heated* serum (containing the antibody which would have hæmolysed them if the complement had not been destroyed) are said to be sensitised, because, if they are introduced into a solution containing the complement, hæmolysis follows immediately.

Not all types of reaction between antibodies and antigens require complement for their completion, but nearly all antigen-antibody compounds have the property of absorbing complement, and thus making it inactive or *fixed*. Complement fixation can be used as a test for the presence of either antibody or antigen. The most famous of these complement-fixation tests is the *Wassermann reaction*, which is a test for the presence of the syphilitic antibody in the blood of a patient suspected of suffering from syphilis. A standard amount of antigen and a standard amount of complement are added to the heated serum of the patient. If the syphilitic antibody is present, it reacts with the antigen ; complement is fixed, and sensitised red cells added subsequently are not hæmolysed. Curiously enough, the standard antigen used in the Wassermann test has nothing to do with syphilis, but is prepared from an alcoholic extract of heart muscle. Presumably this empirical antigen, unexpectedly discovered in the course of controls on the Wassermann reaction, has the same chemical configuration as some antigen in the causal organism of syphilis, for exhaustive tests have only confirmed the usefulness of the reaction in the diagnosis of syphilis.

Anaphylactic Shock. If a single injection of an antigen is followed after an interval of about ten to fourteen days by an injection of a second dose of the same antigen, the consequence is a profound, and often fatal, collapse, due to a very low blood pressure resulting from dilatation and increased permeability of the capillaries ; in some animals, asphyxia is induced by intense constriction of the bronchi. The first dose of the antigen clearly rendered the animal hypersensitive, instead of immune, the hypersensitivity being associated with a *low* content of antibodies in the blood. A widely supported view of the mechanism of anaphylactic shock supposes that the free circulating antibodies due to the first injection are sufficient to " neutralise " only a part of the second dose of antigen ; the remainder of the antigen reacts with antibodies which are attached to tissue cells. This reaction damages the tissue cells in some way, leading to the production of histamine, which is known to produce, when injected intravenously, a train of events very like anaphylactic shock. If the animal in the hypersensitive state is given a series of injections of the antigen, each too small to produce shock, it will, in time, become desensitised, presumably because an adequate supply of circulating antibodies is developed.

On rare occasions a condition resembling anaphylactic shock results from an intravenous injection of an antitoxin. This is usually due to the patient being hypersensitive to the horse serum from which most antitoxins are prepared.

There is a mild type of anaphylactic response, with much less violent

manifestations of hypersensitivity, known as an "allergic reaction." This may follow the consumption of mussels, lobsters, strawberries or several common foodstuffs by certain persons who are said to be "sensitive"; or it may follow contact of the skin or mucous membranes of the respiratory passages with the pollen of certain grasses (in "hay fever"), the hairs of certain animals, or even the close presence of these animals. The nature of the reactions produced depend on the nature of the antigen (or "allergen"), and may vary considerably from one person to another. Characteristic reactions are nettle-rash and urticaria—*i.e.*, the eruption of wheals on the skin; congestion and excessive irritability of the mucous membranes of the nose and pharynx; and constriction of the bronchioles, producing asthma (Chapter 4, p. 97).

Blood Groups

It is well known that the effects of severe loss of blood are usually best countered by transfusion of blood from another individual. Early attempts at such transfusion often had disastrous results, owing to the fact that the injected red cells may clump together (agglutinate) in large masses which block certain of the capillaries in the body (Fig. 21. 1); the cells then hæmolyse, and the liberated hæmoglobin is in part converted to bilirubin, with consequent jaundice, and in part excreted by the kidneys; the secretion of urine is impaired, or may even stop. When such effects follow the transfusion, the blood of the donor is said to be "incompatible" with that of the recipient. This incompatibility was explained when it was discovered that human serum may contain antibodies which act on the red cells of certain other individuals, making them stick together (agglutinins) or break up (lysins); these are termed "naturally occurring" antibodies, since they have not been formed in response to the presence of known antigens. To be susceptible to the agglutinins, the red cells must contain agglutinogens (*i.e.*, antigens) with which the agglutinins react. The experimental facts were found to be explicable by the hypothesis that two kinds of agglutinogen, A and B, are to be found in human red cells. In some bloods the red cells contain agglutinogen A, in others they contain B, in others both A and B together, and in still others they contain neither. Thus blood can be classified into four groups according as their cells contain the agglutinogens, A, B, AB or O. Similarly, it is postulated that there are in human sera two agglutinins, α and β, which react respectively with agglutinogens A and B. Obviously, in any normal blood, the corresponding agglutinins and agglutinogens which would react with each other cannot be present at the same time. Consequently, only in O blood are α and β agglutinins to be found together. In A blood only β agglutinin is present, in B blood only α, while in AB blood neither α nor β is present.

The blood of a particular man can be easily assigned to its proper group if samples of serum from blood of groups A and B are available as is indicated in Fig. 21. 1, which shows the effect of serum from each group on cells of each of the other groups. In blood transfusion it is always desirable to use a donor of the same group as the recipient, but in emergency it is considered allowable to use any donor whose cells

are not agglutinated by the serum of the recipient. The donor's cells are exposed to the full effect of the recipient's serum, whereas the donor's serum is diluted by the greater volume of the recipient's blood, and hence is not likely to harm the recipient's cells. As can be seen in Fig. 21. 1, cells of group 0 are not agglutinated by any type of serum ; people with blood of group O are thus called *universal donors.* Similarly,

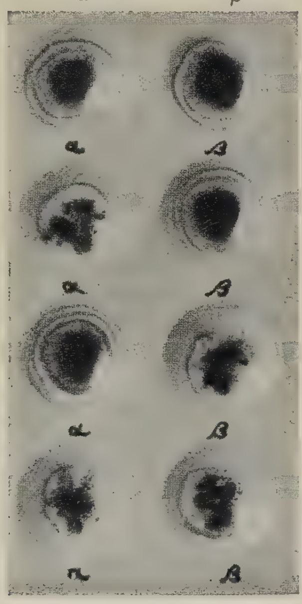


FIG. 21. 1. The **Agglutination of Red Blood Corpuscles,** and the **Four Blood Groups** in Man.

Four large drops of serum of group B (containing agglutinin α) and four large drops of serum of group A (containing agglutinin β) were placed on the slide. To each of the top two drops was added a small drop of blood of group O ; to the second two, a drop of blood of group A ; to the third two, a drop of blood of group B ; and to the bottom two, a drop of blood of group AB. Agglutination only occurred when the agglutinogens on the corpuscles met the corresponding agglutinins in the serum. (Lattes' "Individuality of the Blood.")

serum from group AB will not agglutinate cells of any group ; people with blood of group AB are thus called *universal recipients.*

More detailed and extensive studies of blood groups have shown that the matter is much more elaborate and complicated than was at first thought. Agglutinogen A consists of two parts or varieties A_1 and A_2 ; and there may be agglutinogens on human red cells which may be classified into at least ten groups or systems, in addition to the ABO system, and to which identifying letters and names have been given. These are of little consequence in transfusion, since normal human sera very rarely contain the corresponding agglutinins ; they are responsible, however, for the fact that on rare occasions two bloods are found to be unexpectedly incompatible.

Blood groups are inherited according to Mendelian laws, the presence of each of the agglutinogens on the red cells being decided by the presence of a certain gene in the chromosomes of the individual concerned (compare Chapter 10, p. 300). They may thus be of practical value in proving non-paternity in law suits concerning the paternity of illegitimate children. The chance of getting positive evidence is increased by the fact that all the agglutinogens—those which are of little importance in transfusion as well as those of the ABO group—are inherited, and their presence can be detected by appropriate immunological tests. The blood groups, indeed, are of considerable interest and importance in the study of human genetics.

Different races of people tend to have different blood groups. In the people of western and north-western Europe, for example, groups A and O are the most common, and the proportion of the population who are of group A becomes smaller as we proceed east across Asia ; in Central Asia and India, group B is the most common, while the Indians of Central and South America may be almost entirely of group O. Study of the blood groups, therefore, particularly when those of the systems other than ABO are included, is of great value in anthropology, indicating possible inter-relations between different races of mankind.

The group of agglutinogens known as the *Rh factors* were first discovered on the red cells of Rhesus monkeys (hence the abbreviation), but they also occur on human cells in some individuals (85 per cent. of Europeans). The corresponding agglutinins do not normally occur, even in the 15 per cent. of Rh negative persons. Injection of Rh positive blood into a Rh negative person will, however, lead to the production of the agglutinins. A later transfusion, therefore, of more Rh positive blood will have serious consequences. As with the other agglutinogens, the presence or absence of the Rh factors is inherited. The fœtus within a Rh negative mother may, by inheritance from the father, be Rh positive. The agglutinins are then formed in the mother's blood, pass into that of the fœtus, and lead to destruction of the red cells, and usually miscarriage or death of the infant shortly after birth. The only remedy is to remove the agglutinins by complete replacement of the infant's blood with that from a normal person.

It is thus advisable before transfusion to test directly the donor's cells against the recipient's serum, and the recipient's cells against the donor's serum, a procedure known as the *cross agglutination* test. This is a precaution not only against errors in grouping, but also tests for the presence of the Rh factors and of the other agglutinins which may occasionally be present.

Rouleaux Formation and Sedimentation Velocity

Red blood corpuscles in plasma nearly always show a tendency to come together with their broad surfaces in apposition, thus forming aggregates which look like rolls of coins and hence have been named *rouleaux*. In practice it is never difficult to distinguish these orderly rouleaux of ten or twenty cells from the disorderly clumps of thousands of cells found in the agglutination reactions. The tendency to form rouleaux varies among different individuals and different species, and largely determines the suspension stability of the blood, *i.e.*, the time required for the red cells to sediment down completely in blood (made

incoagulable) which is allowed to stand. This is due to the fact that the rate at which a system of suspended particles settles increases with the size of the particles, other factors remaining constant. The formation of rouleaux may be regarded as a mild, and reversible, kind of agglutination : it occurs more readily when the concentration of fibrinogen or globulin in the plasma is increased, and it may be related to the presence of antibodies and to the occurrence of immunity reactions. An increased sedimentation velocity has been observed to accompany most inflammatory diseases, and also to accompany pregnancy ; its measurement, therefore, is of diagnostic interest.

Sedimentation sometimes occurs so rapidly that a clear layer of plasma is left before coagulation begins ; the clot is thus partly free from corpuscles and forms what is known as the *buffy coat* on the surface of the corpuscular mass. This fact has been known since the days of the Greeks, and very largely formed the basis of the practice of blood-letting as a cure for all diseases. It was thought that the buffy coat was formed by the foul matter in the blood that was responsible for the disease ; the more blood one could remove, therefore, the quicker would the patient recover. Accidents sometimes happened, however, through the physician mistaking the normal effect of pregnancy for evidence of a pathological condition.

Grafting

A piece of skin, or even a whole organ, may be removed from an individual and transferred, or " transplanted," to some other place in the same individual, where it may survive and retain its functions. But if it is transferred to another individual, even of the same species of animal, it will not survive, but, after a week or two, will be destroyed and removed just as if it were a " foreign " substance, and with the same kind of inflammatory reactions. This is due to the formation of antibodies which react with the antigens of the transplanted tissue. If the animal on which the skin, for example, is grafted has been " immunised " (" sensitised " would describe the phenomena better in this instance) by a previous graft from the same animal, or the injection of a suspension of living cells from this animal, the graft will be destroyed very rapidly, the necessary antibodies being already present. But if the donor and recipient of the graft are identical twins, have developed from the same fertilised egg-cell, and have the same genes in their chromosomes, a graft from one will " take " in the other. Studies of this kind lead to the conclusion that no two individuals (apart from identical twins) have cells and body fluids of precisely the same chemical composition, any more than they have precisely the same facial appearance or finger-prints.

One of the consequences of exposure of an animal or man to ionising radiations—from nuclear explosions, or the experimental and industrial use of radio-active isotopes and X-ray apparatus—is the destruction or derangement of the tissues from which both the white and the red blood cells are derived. This leads to serious and sometimes fatal abnor-

malities of the blood cells, as in leukæmia and anæmia (Chapter 3, p. 72). It has been found that the lethal effects may be prevented by injecting suspensions of cells from the blood-forming tissues of another individual; these replace those damaged by the radiation and take over their functions. This has been demonstrated in mice, for example, tissue even from rats having " taken " and rat blood cells found in circulation; it is not certain that it occurs satisfactorily in man. But the fact that such a transplantation " takes," at least in some kinds of animal, suggests that the radiations have also damaged the tissues which are responsible for the production of antibodies—in this instance a fortunate combination of effects.

EPILOGUE

THE functions of the various organs in the body have been described in the preceding chapters with rather special emphasis on the properties of the organs or physiological systems concerned. In living animals and man, however, an organ or tissue rarely increases or decreases its activity without affecting the activities of many other organs or tissues. These are all controlled by the nervous and endocrine systems so as to promote, as far as possible, the well-being and stability of the animal as a whole and the preservation of the species.

To a large extent this control is directed towards maintaining at steady values the volume, chemical composition (content of oxygen, carbon dioxide, glucose, salts, *p*H, etc.) and temperature of the body fluids (the " internal environment "), as already described in several chapters; any or all of these are likely to be disturbed by change of activity in one part of the body and restored by compensatory changes in other parts. It was this kind of regulatory process to which the term " physiological homeostasis " was originally given. The concept, however, may be extended to many aspects of our relations with the external world; for example, maintenance of the characteristic posture and orientation with respect to gravity. These, of course, are not necessarily constant. We are capable of movement and of adjusting our behaviour according to changes in the environment, and in general this involves co-ordinated actions directed towards maintaining continued existence in the " normal " state. If we are threatened by some object in the external world, our muscles are activated for self-preservation by taking evasive or hostile action as instinct, experience and instruction dictate; the consequent disturbance of the internal environment brings into play all the necessary restorative processes. These will include not only those involving, for example, the cardio-vascular and respiratory systems, but also the acquisition, intake, digestion and absorption of food and water to replace the stores used up.

The first step in the maintenance of homeostasis is the detection by appropriate receptors of small changes in the various states and conditions of the external and internal environments. The messages, or " signals ", which they originate pass to various parts of the nervous and endocrine systems where they are modified and co-ordinated and sent out again (" reflected ") to various effectors, *e.g.*, muscles and secreting glands, which can change the position and movement of the body and the physical and chemical states of the internal environment. The changes that result affect the receptors which originated the signals, that is, there is " feed-back " from the effectors to the receptors as well as " reflex " control of the effectors by the receptors. It is an important feature of such a system that the feed-back should be " negative ", so arranged that any change which disturbs the steady state induces a response which acts so as to reduce the disturbance and restore the

initial or " standard " conditions. The actions of the effectors then cease, or become stabilised, only when the receptors cease to originate signals, or originate some standard pattern of signal; they are continuously " monitored " and adjusted so as to conform to a co-ordinated pattern. Movement of a leg, for example, involves much more than the excitation and contraction of certain leg muscles; the nervous system " thinks " in terms of co-ordinated movements so that there is an appropriate and varying excitation of some muscles and inhibition of others, monitored by signals from many kinds of receptor. Likewise, secretion of the trophic hormones of the anterior pituitary gland is reduced or suppressed by the hormones which they cause to be secreted by the target organs; this negative feed-back, from target organ to anterior pituitary gland, keeps the amount of these hormones in the body relatively constant. Homeostasis is maintained, therefore, by the action of systems which are essentially the same as those known to engineers as " servo-systems ".

Imagine a man hunting or reaping for his food supplies; or, a more likely occupation for our readers, imagine him in some active game or sport. His eyes are directed so that the images of the object aimed at lie on the foveæ—a good example of " automatic following " by means of a servo-system. His arm muscles are activated so as to direct, say, the arrow, reaping-hook or implement of sport until the pattern of signals from his eyes reaches the desired standard condition such that he will hit the prey, cut the corn or drive the ball. His leg muscles may be activated so as to produce the movements of running. There is thus a continuous adjustment of contractions of opposing muscles, while balance and posture are maintained by the excitation and inhibition of most of the other muscles in the body; all being controlled by combining information derived from the eyes, the receptor organs for balance (semi-circular canals, etc.) and receptors in the skin, joints and muscles themselves. Unless this control is smooth and accurate, brought about by well co-ordinated activity of many parts of the nervous system, energy will be wasted in checking needless movements and in failure to achieve the goal desired.

The active muscles use oxygen and glucose and produce carbon dioxide and heat more rapidly than they did when at rest. The flow of blood through them is increased and, through the action of the vasomotor control system, that through the temporarily unimportant abdominal organs is decreased in compensation. The temperature regulating centres ensure that the flow through the skin is also increased so that the extra heat generated is dissipated. By the action of the " muscle pump " and the secretion of adrenaline, the output of the heart and flow of blood through the lungs are increased; more oxygen is removed from the lungs and more carbon dioxide delivered to them. The increased amount of carbon dioxide in the body, the decreased amount of oxygen, and other associated departures from the " standard " state, excite the respiratory centres; more air is breathed in and out of the lungs, extra oxygen is supplied and more carbon dioxide carried away.

The fall in blood sugar concentration leads to a release of glucose from the glycogen stores of the liver. Water is lost from the body in the dissipation of heat, particularly if the exercise is severe enough to cause sweating; excitation of the osmo-receptors results in secretion of anti-diuretic hormone by the posterior pituitary gland and the kidneys excrete as little water as possible until the loss is made good by drinking. All these homeostatic processes, moreover, are adjusted continuously, mainly through the autonomic nervous system and secretion of hormones, as conditions change throughout the period of exercise and subsequent recovery. Nearly every organ of the body, therefore, is affected by any form of severe exercise; and, indeed, the exploration of the consequences of exercise on various organs has been one of the most fertile fields of experimental inquiry in physiology.

BIBLIOGRAPHY

The books and review articles listed below have been chosen from the many available for the guidance of students who wish to penetrate more deeply into certain branches of physiology. A few of the larger textbooks have been included for reference purposes. Medical students mostly find them too comprehensive to read from cover to cover, but they are of great value for a more detailed study of the parts of the subject found most interesting. Many of the best accounts of human physiology are to be found in the introductory chapters of monographs primarily addressed to readers with clinical interests, and a number of such monographs have accordingly been included. Many books and articles included in the list represent the best available account of their field although not the whole of their content is up to date. A few have been included more for the sake of their stimulating outlook than for their detailed accuracy. Readers should, therefore, carefully note the dates of publication and bear in mind the possibility that part of the content may have been superseded by more recent discoveries.

Current views on specialised topics will be found in articles in Physiological Reviews, Ergebnisse der Physiologie (some articles in English), Biological Reviews, Pharmacological Reviews, Harvey Lectures, and the Cold Spring Harbor Symposia. Students, in their spare time (if any is permitted to them) should make a habit of consulting the volumes issued within the last ten years, and reading articles on subjects in which they are most interested. For review articles on the most recent developments, and for lists of references to modern papers, students should refer to the current volumes of the Annual Reviews of Physiology, of the Annual Reviews of Biochemistry, of Vitamins and Hormones, and of any other Annual Reviews that may appear.

GENERAL

Principles of Human Physiology, by *E. H. Starling and C. Lovatt Evans.* 13th. edn. ed. by *H. Davson and M. G. Eggleton.* 1962. 1579 pp.

Medical Physiology, ed. by *Philip Bard.* 11th. edn. 1961. 1339 pp.

The Physical Basis of Medical Practice, by *C. H. Best and N. B. Taylor.* 7th. edn. 1961. 1554 pp.

Textbook of General Physiology, by *H. Davson.* 2nd. edn. 1959. 846 pp.

Comparative Animal Physiology, by *C. L. Prosser and F. A. Brown.* 2nd. edn. 1961. 688 pp.

Principles of General Physiology. 5th. edn. by *L. E. Bayliss.* Vol. 1, "The Physico-chemical Background". 1959. 520 pp. Vol. 2, "General Physiology". 1960. 848 pp.

Features in the Architecture of Physiological Function, by *J. Barcroft.* 1934. 368 pp.

The Wisdom of the Body, by *W. B. Cannon.* 1932. 312 pp.

Adventures in Physiology, with Excursions into Autopharmacology, by *Sir Henry Dale.* 1953. 652 pp.

Human Physiology: A Practical Course, by *C. G. Douglas and J. G. Priestley.* 1948. 229 pp.

Mammalian Physiology, by *E. G. T. Liddell and C. S. Sherrington.* 1929. 162 pp.

Experimental Physiology, by *D. T. Harris, H. P. Gilding and W. A. M. Smart.* 1956. 6th. edn. 305 pp.

Electronic Apparatus for Biological Research, by *P. E. K. Donaldson.* 1958. 718 pp.

Dynamic Aspects of Biochemistry, by *E. Baldwin.* 3rd. edn. 1957. 526 pp.

TEXTBOOK OF BIOCHEMISTRY, by *E. S. West and W. R. Todd.* 3rd. edn. 1961. 1423 pp.
PRINCIPLES OF BIOCHEMISTRY, by *A. White, P. Handler, E. L. Smith and De Witt Stetten.* 2nd. edn. 1959. 1149 pp.
AIDS TO BIOCHEMISTRY, by *S. P. Datta and J. H. Ottoway.* 5th. edn. 1960. 266 pp.
PRACTICAL PHYSIOLOGICAL CHEMISTRY, by *P. B. Hawk, B. L. Oser and W. H. Summerson.* 13th. edn. 1954. 1439 pp.
PRACTICAL PHYSIOLOGICAL CHEMISTRY, by *S. W. Cole.* 10th. edn., by *E. Baldwin and D. J. Bell.* 1955. 261 pp.
PHYSICAL CHEMISTRY FOR STUDENTS OF PHARMACY AND BIOLOGY, by *S. C. Wallwork.* 1956. 307 pp.
TEXTBOOK OF HISTOLOGY, by *A. A. Maximow and W. Bloom.* 7th. edn. 1957. 628 pp.
THE PHARMACOLOGICAL BASIS OF THERAPEUTICS, by *L. Goodman and A. Gilman.* 2nd. edn. 1955. 1831 pp.
PHARMACOLOGY, by *J. H. Gaddum.* 5th. edn. 1959. 587 pp.
CLARK'S APPLIED PHARMACOLOGY, by *A. Wilson and H. O. Schild.* 9th. edn. 1959. 750 pp.
HISTORY OF BIOLOGY: A GENERAL INTRODUCTION TO THE STUDY OF LIVING THINGS, by *C. Singer.* 1950. 579 pp.
SHORT HISTORY OF PHYSIOLOGY, by *K. J. Franklin.* 1949. 147 pp.
SELECTED READINGS IN THE HISTORY OF PHYSIOLOGY, by *J. F. Fulton.* 1930. 317 pp.

CIRCULATION, RESPIRATION

HANDBOOK OF PHYSIOLOGY, published by *American Physiological Society.* Section 2, " Circulation ". Vol. 1. 1962. 758 pp.
THE LAW OF THE HEART, by *E. H. Starling. Linacre Lecture.* 1918. 27 pp.
THE MECHANISM AND GRAPHIC REGISTRATION OF THE HEART BEAT, by *Thomas Lewis.* 1925. 452 pp.
THE PRESSURE PULSES OF THE CARDIOVASCULAR SYSTEM, by *C. J. Wiggers.* 1928. 200 pp.
SYMPOSIUM ON THE REGULATION OF THE PERFORMANCE OF THE HEART, by *L. N. Katz and others. Physiol. Rev.*, **35**, 91–168 (1955).
CARDIAC CONTROL, by *R. F. Rushmer and O. A. Smith. Physiol. Rev.*, **39**, 41–68 (1959).
THE BLOOD VESSELS OF THE HUMAN SKIN AND THEIR RESPONSES, by *Thomas Lewis.* 1927. 322 pp.
THE ANATOMY AND PHYSIOLOGY OF CAPILLARIES, by *August Krogh.* 1929. 422 pp.
SYMPATHETIC CONTROL OF HUMAN BLOOD VESSELS, by *H. Barcroft and H. J. C. Swann.* 1952. 165 pp.
VENOUS RETURN, by *G. A. Brecher.* 1956. 148 pp.
REFLEXOGENIC AREAS OF THE CARDIOVASCULAR SYSTEM, by *C. Heymans and E. Neil.* 1958. 271 pp.
SYMPOSIUM ON CENTRAL NERVOUS SYSTEM CONTROL OF CIRCULATION. *Physiol. Rev.*, **40**, Supplement No. 4. 1960. 311 pp.
VOLUME AND DISTRIBUTION OF THE BLOOD AND THEIR SIGNIFICANCE IN REGULATING THE CIRCULATION, by *T. Sjöstrand. Physiol. Rev.*, **33**, 202–228 (1953).
FUNCTIONAL SIGNIFICANCE OF VENOUS BLOOD PRESSURE, by *E. M. Landis and J. C. Hortenstine. Physiol. Rev.*, **30**, 1–32 (1950).
BLOOD: A STUDY IN GENERAL PHYSIOLOGY, by *L. J. Henderson.* 1928. 397 pp.
THE RESPIRATORY FUNCTION OF THE BLOOD, by *J. Barcroft.* 2 vols. 1925, 207 pp.; 1928, 200 pp.
HÆMOGLOBIN, ed. by *F. J. W. Roughton and J. C. Kendrew.* 1949. 317 pp.

Respiration, by *J. S. Haldane and J. G. Priestley.* 1935. 493 pp.
Pulmonary Ventilation and its Physiological Regulation, by *J. S. Gray.* 1950. 82 pp.
Organisation of the Respiratory Centre, by *R. F. Pitts. Physiol. Rev.*, **26**, 600–630 (1946).
Analysis of Factors Concerned in Regulation of Breathing in Exercise, by *F. S. Grodins. Physiol. Rev.*, **30**, 220–239 (1950).
The Lung. Clinical Physiology and Pulmonary Function Tests, by *J. H. Comroe and others.* 1955. 219 pp.

DIGESTION

Human Gastric Function, by *S. Wolf and H. G. Wolff.* 1947. 262 pp.
The Secretory Mechanism of the Digestive Glands, by *B. P. Babkin.* 1950. 1027 pp.
Physiology of the Digestive Tract, by *H. W. Davenport.* 1961. 221 pp.
Secretory Mechanisms of the Gastro-intestinal Tract, by *R. A. Gregory.* 1961. 248 pp.
Physiology of the Salivary Glands, by *A. S. V. Burgen and N. G. Emmelin.* 1961. 279 pp.
The Intestinal Tract, by *R. P. Spencer.* 1960. 411 pp.
Gastric Emptying and Secretion in Man, by *J. N. Hunt. Physiol. Rev.*, **39**, 491–533 (1959).

METABOLISM, NUTRITION

Metabolic Pathways, ed. by *D. M. Greenberg.* 2nd. edn. Vol. 1. 1960. 572 pp. Vol. II. 1961. 814 pp.
Symposium on Metabolism of Lipids, by *G. Popjak and others. Brit. Med. Bulletin*, **14**, 197–273.
Carbohydrate Metabolism, by *S. Soskin and R. Levine.* 1952. 346 pp. [See also Textbooks of Biochemistry.]
The Elements of the Science of Nutrition, by *G. Lusk.* 4th. edn. 1928. 844 pp.
Biochemistry and Physiology of Nutrition, ed. by *G. H. Bourne and G. W. Kidder.* 1953. Vol. I, 569 pp. Vol. II, 641 pp.
Modern Nutrition in Health and Disease, ed. by *M. G. Wohl and R. S. Goodhart.* 2nd. edn. 1960. 1152 pp.
Human Nutrition and Dietetics, by *S. Davidson, A. P. Meiklejohn and R. Passmore.* 1959. 844 pp.
Human Energy Expenditure, by *R. Passmore and J. V. G. A. Durnin. Physiol. Rev.*, **35**, 801–859 (1955).
Thirst, by *A. V. Wolf.* 1958. 536 pp.

BODY FLUIDS, URINE

The Body Fluids, by *J. R. Elkington and T. S. Danowski.* 1955. 626 pp.
Body Water in Man: the Acquisition and Maintenance of the Body Fluids, by *M. B. Strauss.* 1957. 286 pp.
Passage of Molecules through Capillary Walls, by *J. R. Pappenheimer. Physiol. Rev.*, **33**, 387–423 (1953).
The Lymphatic System, by *C. K. Drinker.* Lane Memorial Lecture. 1942. 101 pp.
Lymphatics and Lymphoid Tissue, by *J. M. Yoffey and F. C. Courtice.* 1956. 510 pp.
Physiology of the Ocular and Cerebrospinal Fluids, by *H. Davson.* 1956. 388 pp.
Physiology of the Kidney, by *Homer W. Smith.* 1952. 1049 pp.

THE PHYSIOLOGICAL BASIS OF DIURETIC THERAPY, by *R. F. Pitts.* 1959. 332 pp.

MODERN VIEWS ON THE SECRETION OF URINE, ed. by *F. R. Winton.* The Cushny Memorial Lectures. 1956. 292 pp.

PHYSIOLOGY OF MICTURITION, by *O. R. Langworthy, L. C. Kolb and L. G. Lewis.* 1940. 232 pp.

REPRODUCTION, DUCTLESS GLANDS

PHYSIOLOGY OF REPRODUCTION, by *F. H. A. Marshall.* New edn. ed. by *A. S. Parkes.* Vol. 1.1, 1956, 688 pp.; Vol. 1.2, 1960, 877 pp.; Vol. 2, 1952, 880 pp.

REPRODUCTIVE PHYSIOLOGY, by *A. V. Nalbandov.* 1958. 271 pp.

ESSENTIALS OF HUMAN REPRODUCTION, by *J. T. Velardo.* 1958. 270 pp.

REPRODUCTION AND SEX, by *G. I. M. Swyer.* 1954. 280 pp.

PHYSIOLOGY OF FERTILISATION IN MAMMALS, by *M. Chang and G. Pincus. Physiol. Rev.*, **31**, 1–26 (1951).

THE PHYSIOLOGY AND BIOCHEMISTRY OF LACTATION, by *S. J. Folley.* 1956. 153 pp.

PHYSIOLOGY OF THE MAMMARY GLANDS, by *J. L. Linzell. Physiol. Rev.*, **39**, 534–576 (1959).

THE HORMONES: PHYSIOLOGY, CHEMISTRY AND APPLICATIONS, ed. by *G. Pincus and J. V. Thimann.* Vol. I, 1948, 886 pp.; Vol. II, 1950, 782 pp.

TEXTBOOK OF ENDOCRINOLOGY, ed. by *R. H. Williams.* 3rd. edn. 1962. 1204 pp.

HORMONES IN BLOOD, ed. by *C. H. Gray and A. L. Bacharach.* 1961. 655 pp.

BODILY CHANGES IN FEAR, RAGE AND PAIN, by *W. B. Cannon.* 1929. 404 pp.

SYMPOSIUM ON ADRENERGIC MECHANISMS (*Ciba Foundation*). 1960. 632 pp.

THE ADRENAL CORTEX, by *I. Chester Jones.* 1957. 316 pp.

REGULATION OF ALDOSTERONE SECRETION, by *G. Farrell. Physiol. Rev.*, **38**, 709–728 (1958).

THE THYROID HORMONES, by *R. Pitt-Rivers and J. R. Tata.* 1959. 247 pp.

THE PARATHYROID GLANDS AND METABOLIC BONE DISEASE, by *F. Albright and E. C. Reifenstein Jr.* 1948. 393 pp.

NEURAL CONTROL OF THE PITUITARY GLAND, by *G. W. Harris.* 1955. 298 pp.

THE ANTIDIURETIC HORMONE AND THE FACTORS WHICH DETERMINE ITS RELEASE, by *E. B. Verney.* Croonian Lecture. *Proc. Roy. Soc. Lond.*, **B.135**, 35–105 (1947).

NERVOUS SYSTEM

HANDBOOK OF PHYSIOLOGY, published by *American Physiological Society.* Section 1, " Neurophysiology ". 3 vols. 1959–60. 2013 pp.

ELECTRIC EXCITATION OF NERVE, by *B. Katz.* 1939. 151 pp.

THE ELECTRICAL SIGNS OF NERVOUS ACTIVITY, by *J. Erlanger and H. S. Gasser.* 1937. 221 pp.

THE IONIC BASIS OF ELECTRICAL ACTIVITY IN NERVE AND MUSCLE, by *A. L. Hodgkin. Biol. Rev.*, **26**, 339–409 (1951).

IONIC MOVEMENTS AND ELECTRICAL ACTIVITY IN GIANT NERVE FIBRES, by *A. L. Hodgkin.* Croonian Lecture. *Proc. Roy. Soc. Lond.*, **B.148**, 1–37 (1958).

THE TRANSMISSION OF IMPULSES FROM NERVE TO MUSCLE AND THE SUBCELLULAR UNITS OF SYNAPTIC ACTION, by *B. Katz.* Croonian Lecture. *Proc. Roy. Soc. Lond.*, **B.155**, 455–477 (1962).

GENERAL PRINCIPLES OF NERVOUS ACTIVITY, by *E. D. Adrian. Brain*, **70**, 1–17 (1947).

THE PHYSIOLOGY OF NERVE CELLS, by *J. C. Eccles.* 1957. 270 pp.

The Integrative Action of the Nervous System, by *C. S. Sherrington.* 1906. 411 pp.; reprinted 1947, 433 pp.
The Selected Writings of Sir Charles Sherrington, ed. by *D. Denny-Brown.* 1939. 532 pp.
Reflex Activity of the Spinal Cord, by *R. S. Creed, D. Denny-Brown, J. C. Eccles and C. S. Sherrington.* 1932. 183 pp.
Anatomy of the Nervous System, by *S. W. Ranson*, revised by *S. L. Clark.* 10th. edn. 1959. 581 pp.
Physiology of the Nervous System, by *J. F. Fulton.* 3rd. edn. 1949. 667 pp.
Physiology of the Nervous System, by *E. G. Walsh.* 1957. 563 pp.
Introduction to Clinical Neurology, by *Gordon Holmes.* 1952. 189 pp.
Critical Studies in Neurology, by *F. M. R. Walshe.* 1948. 256 pp.
Selected Papers of Sir Gordon Holmes, ed. by *F. M. R. Walshe.* 1956. 262 pp.
The Functional Organisation of the Diencephalon, by *W. R. Hess.* 1957. 180 pp.
The Cerebellum of Man, by *Gordon Holmes. Brain*, **62**, 1–30 (1939).
The Basal Ganglia, by *D. Denny-Brown.* 1962. 144 pp.
The Cerebral Cortex of Man, by *W. Penfield and T. Rasmussen.* 1950. 248 pp.
The Parietal Lobes, by *Macdonald Crutchley.* 1953. 480 pp.
The Physical Background of Perception, by *E. D. Adrian.* 1947. 95 pp.
Receptors and Sensory Perception, by *R. Granit.* 1959. 369 pp.
Nervous Gradation of Muscular Contraction, by *P. H. Hammond, P. A. Merton and G. G. Sutton. Brit. Med. Bulletin*, **12**, 214–218 (1956).
Conditioned Reflexes, by *I. P. Pavlov*, translated by *G. V. Anrep.* 1927. 430 pp.
The Autonomic Nervous System, by *J. C. White, R. H. Smithwick and F. A. Simeone.* 3rd. edn. 1952. 569 pp.
The Autonomic Nervous System, by *A. Kuntz.* 1954. 605 pp.
Autonomic Neuro-effector Systems, by *W. B. Cannon and A. Rosenblueth.* 1937. 229 pp.

MUSCLES

Living Machinery, by *A. V. Hill.* 1927. 256 pp.
The Structure and Function of Muscle, ed. by *G. H. Bourne.* 1960. Vol. 1. "Structure". 472 pp. Vol. 2. "Biochemistry and Physiology". 593 pp. Vol. 3. "Pharmacology and Disease". 489 pp.
Physiology of Voluntary Muscle, ed. by *W. D. M. Paton. Brit. Med. Bull.*, **12**, 161–236 (1956).
Vertebrate Smooth Muscle, by *E. Fischer. Physiol. Rev.*, **24**, 467–490 (1944).
Conduction, Automaticity and Tonus of Visceral Muscle, by *E. Bozler. Experientia*, **4**, 213–218 (1948).

RECEPTORS, SENSATIONS

The Basis of Sensation, by *E. D. Adrian.* 1928. 122 pp.
Receptors and Sensory Perception, by *R. Granit.* 1955. 369 pp.
Some Principles of Receptor Action, by *H. Davies. Physiol. Rev.*, **41**, 391–416 (1961).
The Chemical Senses, by *R. W. Moncrieff.* 2nd. edn. 1951. 538 pp.
The Physiology of the Eye, by *H. Davson.* 1949. 451 pp.
Vision and the Eye, by *M. Pirenne.* 1948. 187 pp.
The Vertebrate Eye and its Adaptive Radiations, by *G. L. Walls.* 1942. 785 pp.

PHYSIOLOGY OF THE RETINA AND THE VISUAL PATHWAY, by *G. S. Brindley.* 1960. 298 pp.

PHYSIOLOGICAL MECHANISMS OF VISION AND THE QUANTUM NATURE OF LIGHT, by *M. Pirenne. Biol. Rev.*, **31**, 194–241 (1956).

THE VISUAL PIGMENTS, by *H. J. A. Dartnall.* 1957. 216 pp.

SPEECH AND HEARING IN COMMUNICATIONS, by *H. Fletcher.* 1953. 461 pp.

BIOPHYSICS AND PHYSIOLOGY OF THE INNER EAR, by *H. Davis. Physiol. Rev.*, **37**, 1–49 (1957).

THE PHYSIOLOGY OF HEARING, by *I. C. Whitfield. Progress in Biophysics*, **8**, 2–47 (1957).

TEMPERATURE REGULATION

PHYSIOLOGY OF HEAT REGULATION AND THE SCIENCE OF CLOTHING, ed. by *L. H. Newburgh.* 1949. 457 pp.

PHYSIOLOGY OF TEMPERATURE REGULATION, by *J. D. Hardy. Physiol. Rev.*, **41**, 521–606 (1961).

PHYSIOLOGY OF MAN IN THE DESERT, by *E. F. Adolph.* 1947. 357 pp.

MAN IN A COLD ENVIRONMENT, by *A. C. Burton and O. G. Edholm.* 1955. 273 pp.

THE SWEAT GLANDS, by *J. S. Weiner and K. Hellmann. Biol. Rev.*, **35**, 141–186 (1960).

REACTIONS TO INJURY

HUMAN BLOOD COAGULATION AND ITS DISORDER, by *R. Biggs and R. G. Macfarlane.* 3rd. edn. 1962. 560 pp.

BLOOD GROUPS IN MAN, by *R. R. Race and R. Sanger.* 4th. edn. 1962. 424 pp.

FUNDAMENTALS OF IMMUNOLOGY, by *W. C. Boyd.* 1956. 776 pp.

THE HOMOGRAFT REACTION, by *P. B. Medawar.* Croonian Lecture. *Proc. Roy. Soc. Lond.*, **B.149**, 145–166 (1958).

INDEX

Figures in **heavy type** refer to pages in which illustrations or tables occur.

PRINTED IN GREAT BRITAIN BY THE WHITEFRIARS PRESS LTD.
LONDON AND TONBRIDGE

p. 443.